Empirical Essays in Health and Education Economics

Inaugural-Dissertation
zur Erlangung des Grades
Doctor oeconomiae publicae (Dr. oec. publ.)
an der Ludwig-Maximilians-Universität München

2010

vorgelegt von
Amelie Catherine Wuppermann

Referent:	Prof. Dr. Joachim Winter
Korreferent:	Prof. Dr. Florian Heiss
Promotionsabschlussberatung:	1. Juni 2011

Acknowledgements

First and foremost, I would like to thank my supervisor, Joachim Winter. I am grateful for his critical questions, constructive comments and general support that have guided me while writing this dissertation and beyond.

I would further like to thank my co-supervisor, Florian Heiss, for his guidance and continuous encouragement during the past years. Thanks also go to Ludger Wößmann for advice on the fourth chapter of this dissertation and for agreeing to serve on my committee.

In addition, I am indebted to many people for their help with writing this dissertation. In particular, I would like to thank my coauthors, Helmut Farbmacher and Guido Schwerdt, for the collaboration that has been and continues to be a source of inspiration and motivation. I would further like to thank Luc Bissonnette, Nicolas Sauter, Beatrice Scheubel and Hannes Schwandt for valuable suggestions and comments and Peter Ihle and Ingrid Schubert for data support.

I would like to thank Klaus Schmidt for convincing me to spend an academic year at the University of Wisconsin-Madison and my professors and fellow-students at the economics department in Madison for the challenging and inspiring academic experience. I would also like to thank Dana Goldman and the Bing Center for Health Economics for giving me the opportunity to write part of this dissertation at RAND. Without my stays in Madison and at the RAND Corporation in Santa Monica this dissertation would not have been the same.

Finally, I gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft through GRK 801 and from the German Academic Exchange Service.

Contents

List of Tables

List of Figures

Preface

Health and education are two fields of economic research that are of relevance to the political agenda in most developed countries. Governments play a major role in organizing and financing health care and the education system. Market imperfections in both areas are put forward as rationale for government intervention (Poterba, 1996).

Imperfections in the health care market result mostly from informational asymmetries. As future health care needs are to a large extent uncertain, a high proportion of individuals demand insurance. Asymmetric distributions of information between the insurer and the user of health care, however, complicate the organization of health insurance. In particular, adverse selection and moral hazard may occur.

When the individuals who buy insurance are heterogeneous in terms of the insured risk and when they have more information on their risk than the insurer, adverse selection may result in market failure. Given a choice between different insurance products, high risk individuals who know about their risk tend to choose generous coverage. Correspondingly, individuals who know that their risk is lower tend to select out of the respective insurance plans. Lower risk individuals are thus not able to obtain comprehensive insurance coverage at a price corresponding to their risk (Cutler and Zeckhauser, 2000, p. 606ff).

Moral hazard, on the other hand, induces individuals to demand more care than they would if they were not insured. Once individuals are insured, they do not have to bear the full costs of medical treatments. If the insurer cannot observe the individuals' actions or has limited means of judging whether a certain treatment is medically indicated, individuals might demand more – or more expensive – care than necessary. Copayments and other restrictions in the scope of insurance coverage are instruments to counteract this problem (Cutler and Zeckhauser, 2000, p. 576ff).

In the education system, market imperfections are mostly linked to externalities (Poterba, 1996). Not only the individual schooled but also the overall economy might profit from the individual's education. Education might, for example, help to adopt new technologies of production and foster economic growth. When individuals choose their level of education solely based on their individual costs and benefits, and do not take

into account the benefits that accrue for society as a whole, they may choose a lower than socially optimal level of education. It is thus frequently seen as a government responsibility to subsidize the education system in order to induce socially optimal levels of education (Hanushek, 2002, p. 2064f.).

Health and education are on the political agenda today for different reasons. Health care costs have risen over the last decades. As a lot of the spending is public expenditure, reforms of the health care system are needed to reduce costs and to comply with fiscal restrictions. Improving the education system is seen as a key to ensuring international competitiveness. While the question is not finally settled, the evidence is accruing that higher levels of education foster economic growth (Hanushek and Woessmann, 2009).

In order to provide guidance to respective reform efforts, empirical research brings together theoretical concepts, statistical methods and data. This dissertation contributes four empirical studies on questions related to health and education. Each of the studies constitutes a self-contained chapter. The first three chapters focus on the individual assessments of health risks and the organization of health care markets. Chapter 4 examines the impact of teaching methods on student achievement.

The analysis in chapter 1 is motivated by a puzzle in the literature on adverse selection. There is differential evidence of adverse selection in markets that are related to similar risks. In particular, empirical studies on life insurance markets find little evidence of adverse selection but there is evidence of adverse selection in annuity markets. Similarly, for some health insurance markets there is evidence of adverse selection, while others do not seem to be adversely selected. Cohen and Siegelman (2010) suggest that different degrees of private information in the different markets result in different scopes for adverse selection and might thus reconcile the mixed findings. Chapter 1 provides an empirical evaluation of this explanation.

As similar risks are insured in the different markets, individual information on the risks is likely the same. Differences in private information could result from differences in what the insurers know about the risks. The latter difference in turn can stem from differences in information that the insurers collect and use for insurance underwriting, i.e. for risk assessment and calculation of premia.

A challenge of this analysis is thus to find a data set that includes information on individuals' and insurers' risk assessments for potential buyers of insurance. As in many circumstances everyone can potentially buy insurance, the analysis requires a sample of the general population. Within general population surveys, however, typically only a limited set of information on individual health risk factors are available.

The data set used in chapter 1 is an exception. It comes from the English Longitudinal Study of Ageing (ELSA). ELSA collects a large set of variables that are typically also collected in the application process in some insurance markets. In particular, ELSA is one of the rare data sets that includes objectively measured health information collected by a nurse. As applicants in health-related insurance markets, such as private health insurance and life insurance, typically have to disclose information from medical examinations, ELSA is better suited to mimic the insurers' information than other surveys of the general population.

ELSA also comprises variables that can proxy for the potential buyers' information on their risk. The main analysis in chapter 1 uses individuals' assessment of their general health as measure for the individual information. The measure is derived from a question in which individuals are asked to rate their general health on a 5-point scale from very good to very bad. The advantage that this measure has for the purpose of my study is that it is subjective. It cannot be verified by a third party. Information that is only contained in this self-rated health measure therefore necessarily remains private.

Furthermore, ELSA is a panel study and thus tracks individuals over time. This allows to observe whether individuals actually experience the insured events in the future. In order to detect private information the statistical analysis conducted in chapter 1 combines the information on the occurrence of the insured event in the future, the proxy for the potential buyer's information on the risk and the application information that insurers use for risk assessment. It is interpreted as evidence for private information if the proxy for the potential buyer's information helps to predict the future occurrence of the risks when the insurer's information is controlled for.

The analysis shows that depending on the information that is used by the insurer to assess the potential buyers' risk there are varying degrees of private information. The results can help to reconcile the mixed evidence on adverse selection in life insurance and annuity markets and in different health insurance markets. Furthermore, the results indicate that adverse selection may be overcome by accurate risk assessments by the insurers in the different insurance markets.

In contrast to chapter 1, the analysis in chapter 2 takes a direct approach to measuring individual risk perception by asking individuals how they rate their chances of developing different diseases in the future. Following the insights on measuring expectations summarized by Manski (2004), Joachim Winter and I implemented a survey on health expectations in the American Life Panel (ALP), an Internet panel administered by RAND. We elicit the subjective health expectations as numerical probabilities. This

has an advantage over qualitative questions as it allows to gauge the accuracy of the individual information by directly comparing it to objective probabilities.

The analysis presented in chapter 2 uses the subjective probabilities to shed light on the question whether individuals who are overweight or obese are misinformed about their health risks. The accuracy of risk perception among smokers has been analyzed based on similar subjective probabilities by Schoenbaum (1997) and Khwaja et al. (2009). For individuals with excess weight a comparable study is not available so far.

Similar to smoking, obesity is responsible for a large share of health care costs (Cawley and Meyerhoefer, 2010). Furthermore, being obese is at least partly influenced by individual behavior. Behavioral changes could help to decrease an individual's body weight. However, changes in behavior require an effort from the individual. From a rational choice perspective, rational individuals would only engage in behavioral changes if the benefits of doing so exceed the costs. If individuals who are overweight or obese underestimate their health risks, this implies that they underestimate the benefits of behavioral changes.

While smokers are generally not found to underestimate their health risks, the results for the overweight and obese presented in chapter 2 indicate that obese individuals underestimate the risks of certain diseases. These results indicate that there is room for health education programs targeted at the obese population. More drastic campaigns on the risks of obesity similar to the campaigns on the risks of smoking might be needed in order to make individuals aware of their risks.

The third chapter is based on joint work with Helmut Farbmacher. Its focus lies on evaluating the effects of a recent reform in the German statutory health insurance system that increased copayments for doctor visits and prescription drugs in an effort to counteract moral hazard and thereby achieve cost containment. Augurzky et al. (2006), Schreyögg and Grabka (2010), Farbmacher (2010) and Rückert et al. (2008) have analyzed how the reform affected health care use among the statutorily insured in Germany. The studies present inconclusive results on whether the reform changed individual behavior.

The analysis presented in the third chapter contributes to this literature by using a new data set. While the earlier studies rely on survey data, the data analyzed in chapter 3 is insurance claims data. It thus allows to reliably observe health care use. More importantly, however, the analysis goes beyond studying the average effect of the reform. It focuses on the question whether different groups of individuals react differently to the reform. Bago d'Uva (2006) and Winkelmann (2006) find that the health care demand among individuals with high use of health care is less sensitive to

changes in prices than the demand among individuals with low use. We expand on these two studies by changing the focus from demand for health care in general, to the demand for different types of health care, namely visiting two different types of doctors – general practitioners (GPs) and specialists.

In order to conduct this analysis we develop a finite mixture bivariate probit model. On the one hand, this model allows for unobserved heterogeneity in terms of different latent classes of individuals. In particular, we assume that there are two groups of individuals that may differ in any observed and unobserved aspect. However, we do not observe which individual belongs to which group. After the estimation of the model we can characterize the different groups in terms of their health care use. On the other hand, the bivariate probit component of the model allows to estimate how the probability to visit different types of doctors is affected within each of the two latent groups.

Our results indicate that the reform had an effect on individual behavior. On average individuals are 3 percentage points less likely to visit any type of doctor at least once within a year after the reform. In the application of the finite mixture bivariate probit model, we find that the two latent groups can be characterized as being likely to see a doctor and as being less-likely to see a doctor before the reform. The estimated probability of belonging to the likely users is 74%, indicating that this is the majority of the population. Similar to Bago d'Uva (2006) and Winkelmann (2006) we find that the demand for health care is relatively more elastic among individuals who are less-likely to see a doctor before the reform.

Furthermore, distinguishing between the two types of doctor visits, we find that the patterns of use change differently in the different groups. While the reform induces the less-likely users to reduce their probability of visiting both types of doctors, the likely users still visit GPs with the same probability after the reform. However, they visit a specialist with a smaller probability.

This result has an important implication for designing health care systems. The copayments do not force individuals to stop visiting specialists. A large part of the population, however, visits a specialist with a smaller probability and instead focuses on GP visits as consequence of the reform. The GPs' function in the health care system was therefore strengthened by the reform. Even though the reform did not implement gatekeepers directly, meaning that individuals still have a choice between general practice and specialist care, it might have implemented effective incentives that induce a large fraction of the population to use GPs as if they were gatekeepers. As it has been suggested that gatekeeping leads to lower health care costs (Scott, 2000),

the particular reform in the German health insurance might guide a way to introduce gatekeeping without force.

The final chapter draws on joint work with Guido Schwerdt. It shifts the focus from health care reform to improving the education system. It attempts to answer the question whether traditional teaching methods are related to worse educational outcomes than more modern approaches.

We use a data set for 8th grade students in the US from the Trends in International Mathematics and Science Study (TIMSS) in 2003. This data set has two major advantages for our analysis and one disadvantage. The first advantage is that the data allows to match students to teachers and we thus know which student was taught by which teachers. The second advantage is the panel dimension, namely that we observe each student in two subjects. We observe how a student performed on the TIMSS tests in math and in science and which methods the teacher used in teaching the specific subject. Taking differences between math and science for each student we employ a first-difference method that allows us to control for any unobserved student traits that are constant between the two subjects.

The disadvantage of the data set is that there is no random variation in teaching methods. Teachers instead choose freely which methods they employ in teaching. Even though we can control for unobserved student characteristics, such as innate ability or motivation, we are not able to identify causal effects of teaching methods. For example, there might be some unobserved teacher characteristic that induces teachers to choose a specific teaching method. If this teacher characteristic also makes a teacher a good teacher, the estimated effect of the teaching method will be confounded with the effect of the teacher characteristic. We evaluate how important this selection based on unobservable teacher characteristics is by borrowing a method pioneered in Altonji et al. (2005). This method essentially allows to estimate how large selection on unobservable characteristics would have to be compared to selection on observable characteristics in order to explain an entire estimated effect.

The specific focus of our paper lies on comparing teaching using lecture style presentation with teaching based on in-class problem solving. Our results indicate that spending more time in class on lecture style teaching is significantly positively related to student performance, when controlling for subject-constant student traits. However, the application of the method pioneered by Altonji et al. indicates that selection on unobservable teacher traits could explain the entire estimated effect. We therefore refrain from interpreting the estimated effect as causal and do not draw conclusions that call for more lecture style teaching. Nevertheless, based on our results it is implausible

that there is a negative effect of lecture style teaching. As our method helps to control for all other confounding factors, a true negative effect would only be consistent with finding a positive correlation between lecture style teaching and student achievement if good teachers choose to teach with the bad teaching method. This, however, seems to be rather implausible.

Chapter 1

Do I know more than my body can tell? An Empirical Analysis of Private Information in Health-Related Insurance Markets

1.1 Introduction

Information asymmetries about risk types in competitive insurance markets are known to induce inefficient outcomes due to adverse selection. In particular, the seminal model by Rothschild and Stiglitz (1976) and various extensions robustly predict that higher risk individuals buy more insurance coverage than lower risk individuals (Chiappori et al., 2006). The empirical evidence on this matter, however, varies for different insurance markets (Cohen and Siegelman, 2010).

Several explanations have been put forward to reconcile theoretical and empirical findings. Multiple dimensions of asymmetric information, for example private information on risk preferences in addition to private information on risk type, can explain the absence of adverse selection in some insurance markets (De Meza and Webb, 2001; Cutler et al., 2008a). Cohen and Siegelman (2010) suggest the absence of useful private information and thus the absence of scope for adverse selection in some but not in other insurance markets as an additional explanation for the mixed findings.

This chapter evaluates empirically whether different degrees of private information exist in health-related insurance markets, in particular in the markets for life insurance, annuities, and health insurance. It attempts to reconcile the mixed findings that have been presented in the literature on adverse selection in the different insurance markets.

While life insurance markets and private individual health insurance markets do not show evidence of adverse selection, annuity markets and group health insurance markets appear to be adversely selected.

Useful private information on the risk that is to be insured exists if a potential buyer of an insurance product has more information on the risk than the insurer. Even though the markets for life insurance and annuities, and different health insurance markets insure similar risks and individuals thus likely have similar information a priori, differences in private information between these markets could exist, because insurance companies in the different markets collect and use different types of information for insurance underwriting, i.e. for risk classification and the calculation of premia. The variation in information used is at least partly due to legal restrictions. US federal law, for example, prohibits employers from charging premia for health insurance based on health-related information (GAO, 2003). Similarly, several US states have introduced community rating in combination with guaranteed issue laws for individual health insurance, entitling individuals to buy insurance coverage at a price depending exclusively on abstract criteria such as sex, age or geographic region (Lo Sasso and Lurie, 2009). In other private health insurance markets on the contrary health-related information is used in underwriting.[1]

Using data from the English Longitudinal Study of Ageing (ELSA), I analyze the scope for private information on the insured risks in life insurance markets, annuity markets, and health insurance markets by exploiting the different information used in underwriting in the different insurance markets. As ELSA is a panel study, it allows to track individuals over time. I can therefore observe whether events happen to the individuals in the future that would result in insurance claims if the individuals were insured. I call this the information on the future 'realization of risk'. In the main analysis, I interpret it as evidence for private information if a measure of individual's self-rated health (SRH) helps to predict the future realization of risk when all information that the insurer uses in underwriting is controlled for.

SRH is thus employed as a proxy for individual information on the insured risks. It was chosen for two reasons: First, it significantly predicts future health events, like death and the onset of specific health conditions.[2] As individuals with worse SRH are

[1] In addition to legal restrictions, political economy concerns (Finkelstein and Poterba, 2006), specificities of the market structure (Kesternich and Schumacher, 2009), or different demand for underwriting in the different markets (Browne and Kamiya, 2009) could explain the use of different information in underwriting.

[2] For overviews on studies analyzing the relationship between SRH and subsequent death see Idler and Benyamini (1997), and DeSalvo et al. (2006). Banks et al. (2007) provide evidence for relationships between SRH and future diagnoses of diseases.

more likely to die sooner and more likely to get certain diseases in the future, SRH presumably captures useful information on the insured risks in the insurance markets that I focus on. Second, SRH is a non-verifiable measure in the sense that insurance companies have no means to verify whether an individual's statement of SRH is true. This is in contrast to other self-reported measures like an individual's co-morbidities or family health history which can be verified by going back to health records. Its non-verifiability makes SRH particularly valuable for analyzing the existence of private information. Information about health-related risks that is only contained in SRH necessarily remains private.

The ELSA data set contains a broad range of health measures that correspond to the information collected and used by insurance companies in underwriting. ELSA is one of the few longitudinal data sets that provides health data which are objectively measured and reported by a nurse: Results of a blood sample analysis, blood pressure measurement, objectively measured body mass index (BMI), and waist-hip-ratio are available. As the objectively measured data are already available in an early wave of the survey, up to 10 years of follow-up can be used for the analysis.[3]

The realization of the risk that is insured in life insurance and annuity markets is measured by an indicator for whether an individual dies within 10 years after the baseline interview. The realization of the insured risk in health insurance markets is captured by a variable that indicates whether an individual is newly diagnosed or has a recurrence of heart disease, cancer or stroke within 8 years after the initial interview. These conditions belong to the most costly conditions at the per capita level (Druss et al., 2002) and thus seem to be reasonable proxies for the risk insured in health insurance markets, namely high medical expenses.

My results indicate that SRH contains information on dying within or surviving the next 10 years and on being diagnosed with one of the costly major health conditions in the next 8 years, when only a limited number of additional control variables is included in the analysis. With the inclusion of medical information and in particular with the inclusion of the objectively measured health data, however, the explanatory power of SRH is significantly reduced. The explanatory power of SRH for the onset of the costly health conditions even vanishes completely. These results are robust to using different proxies for the individual information on risks.

The results in combination with different use of information for underwriting in different insurance markets can help to explain the mixed evidence of adverse selection in

[3]The US Health and Retirement Study has recently also started to collect objectively measured health information. Up to now at most 1 wave, i.e. 2 years, of follow-up is available, however.

life insurance and annuity markets, and in different health insurance markets. While in life insurance markets stringent underwriting is performed and no evidence is found for adverse selection, in annuity markets only limited information is used for underwriting and evidence for adverse selection is found. Similarly, in individual health insurance with stringent underwriting no evidence is found for adverse selection, while in group health insurance there is evidence for adverse selection and no individual underwriting is performed.

An implication of the results is that the scope for adverse selection in group health insurance markets and annuity markets could be reduced by employing more stringent underwriting. Whether more stringent underwriting would increase welfare, however, cannot be determined based on my analysis. While stringent underwriting might mitigate the effects of adverse selection, it might have consequences not taken into account in my analysis. The overall welfare effect of stringent underwriting depends on the exact institutional setting. Health insurance, for example, is often bought in short-term contracts, that is individuals have to buy coverage repeatedly over time. If every time when individuals buy coverage stringent underwriting is employed, individuals will face the risk of 'becoming a higher risk' and thus having to pay higher prices for insurance in the future. Without an insurance for this additional risk stringent underwriting could reduce welfare.

The chapter is structured as follows: Section 2 gives an overview on the empirical evidence of adverse selection in health-related insurance markets, Section 3 introduces the data used for the analysis. Section 4 outlines the estimation strategy, results are shown in Section 5, Section 6 reports the results of several robustness analyses, and the last section concludes.

1.2 Empirical Literature on Adverse Selection

The empirical literature has analyzed whether adverse selection is present in many different insurance markets, such as the markets for automobile insurance, crop insurance, long term care, reverse mortgages, life insurance, annuities, and health insurance. Cohen and Siegelman (2010) provide a recent overview on the empirical findings. In this section, I focus on the literature concerning the last three markets.

The empirical evidence of adverse selection varies between life insurance and annuity markets. While in markets for annuities evidence for adverse selection is found (Finkelstein and Poterba, 2002, 2004, 2006; McCarthy and Mitchell, 2010; Einav et al., 2010b), there is only very little evidence for adverse selection in life insurance markets.

He (2009) finds evidence that individuals in the Health and Retirement Study (HRS) who newly buy life insurance die earlier than individuals who do not buy life insurance when controlling for variables that insurance companies use for risk classification and calculation of premia. However, as there is no objectively measured health information in the HRS waves that He uses in her analysis, her study might not sufficiently control for risk classification undertaken by the insurers. Furthermore, Cawley and Philipson (1999), Hendel and Lizzeri (2003) and McCarthy and Mitchell (2010) find no evidence for adverse selection in life insurance markets.

With respect to markets for health insurance there are mixed findings on adverse selection in the literature. Especially the US employer-sponsored health insurance market is found to suffer from adverse selection. Cutler and Zeckhauser (2000) provide a summary of these studies. Bundorf et al. (2008), Cutler et al. (2009) and Einav et al. (2010a) provide additional evidence of adverse selection in employer-sponsored health insurance for different US employers. In the non-group health insurance market there is evidence for adverse selection in states where community rating in combination with guaranteed issue laws prohibit health-related underwriting and entitle individuals to buy insurance coverage (Lo Sasso and Lurie, 2009).

Other health insurance markets, however, are typically not found to suffer from adverse selection. Buchmueller et al. (2004) show that sicker individuals are not more likely to buy private supplementary health insurance in France; Propper (1989) and Doiron et al. (2008) find similar results for the British and Australian private health insurance markets. Furthermore, the US market for Medigap coverage is not found to be affected by adverse selection (Fang et al., 2008).

Cutler et al. (2008a) analyze to what degree heterogeneities in risk preferences can explain the mixed evidence of adverse selection in different insurance markets. They find that individuals who engage in risky behavior or refrain from risk lowering activities are less likely to hold life insurance, health insurance and annuities. Refraining from risky behavior and engaging in risk lowering activities are related to prolonging life, and thus to being of lower risk in life insurance and of higher risk in annuity markets. The authors conclude that adverse selection in life insurance markets might be outweighed by selection on risk preferences, while adverse selection in annuity markets is aggravated by it. The mixed evidence for adverse selection in different health insurance markets, however, is difficult to reconcile based on unobserved risk preferences.

The study that is most closely related to mine in terms of the empirical strategy is Banks et al. (2007). The authors use SRH as a proxy for individual information on different future health events to detect scope for adverse selection. The focus of

their study, however, lies in detecting whether there are differential amounts of private information for individuals at different ages. They find that SRH contains more information for older individuals. As objectively measured health data is not included in their data set, the authors cannot fully investigate the effects of medical underwriting on the degree of private information.

1.3 Data

The English Longitudinal Study of Ageing (ELSA) is a rich panel data set which contains socio-demographic, economic, and health-related data for individuals that were born on or before February 29th, 1952 and were living in private homes in England at the time of the interview in 2002. In addition to those core sample members, younger partners living in the same household are interviewed as part of ELSA. The sample was randomly selected from the English population in three repeated cross sections for the Health Survey for England (HSE) in the years 1998, 1999 and 2001. In addition to the data from the eligible ELSA subsample of the three HSE years, called ELSA wave 0, I use data from three ELSA waves collected in 2002, 2004 and 2006.[4]

While ELSA was highly influenced by and modeled on the US Health and Retirement Study (HRS), its design differs from that of the HRS in one important feature: In addition to the biannual interview, every four years a nurse visit is conducted as part of ELSA. Due to these nurse visits objectively measured blood pressure, results of a blood sample analysis, and anthropometric data are available. In wave 0, a blood sample analysis was only carried out for individuals in the 1998 HSE year. As a focus of this study is the scope for adverse selection when insurance companies are allowed to collect and use outcomes of medical screening, the analysis is conducted using only ELSA sample members that were sampled for the 1998 HSE.

The ELSA data in wave 0 contains 8,267 individuals from HSE 1998 (7,807 core sample members and 459 younger partners). Everyone who was interviewed was eligible for the nurse visit in the HSE. Core sample members and younger partners from wave 0 are therefore included in this analysis. As table 1.1 shows, the data contains information on age, gender, race, social occupational class, martial status, and smoking status for nearly all individuals in the sample. Individuals older than 90 are not included in the analysis as their exact age is not known. This reduces the sample size by 1 percent and has no significant effects on the results.

Of particular importance for my analysis is the SRH measure in ELSA. Individuals

[4]For a more thorough description of ELSA see Marmot et al. (2009).

Table 1.1: Descriptives – ELSA Wave 0

	Women			Men		
	1	2	3	4	5	6
Variable	Mean	(% missing)	Mean	Mean	(% missing)	Mean
Age						
Age < 50	.2	(0.00)	.23	.15	(0.00)	.18
Age 50-59	.29	(0.00)	.34	.31	(0.00)	.34
Age 60-69	.24	(0.00)	.25	.27	(0.00)	.30
Age 70-79	.19	(0.00)	.14	.19	(0.00)	.15
Age 80-89	.09	(0.00)	.04	.07	(0.00)	.04
Occupational Class						
Professional	.05	(0.00)	.07	.07	(0.00)	.09
Managerial - technical	.29	(0.00)	.32	.31	(0.00)	.33
Skilled - non manual	.16	(0.00)	.14	.09	(0.00)	.09
Skilled - manual	.25	(0.00)	.27	.33	(0.00)	.32
Semi-skilled manual	.16	(0.00)	.14	.14	(0.00)	.13
Unskilled manual	.07	(0.00)	.05	.05	(0.00)	.04
Other social class	.03	(0.00)	.01	.01	(0.00)	.01
Marital Status						
Married	.62	(0.00)	.70	.77	(0.00)	.80
Widowed	.22	(0.00)	.15	.08	(0.00)	.06
Separated/divorced/single	.16	(0.00)	.16	.15	(0.00)	.14
Activity Status						
Retired	.38	(0.00)	.32	.41	(0.00)	.36
Unemployed	.01	(0.00)	.02	.03	(0.00)	.02
Sick	.04	(0.00)	.02	.08	(0.00)	.06
Working	.36	(0.00)	.45	.47	(0.00)	.54
Inactive	.21	(0.00)	.19	.02	(0.00)	.01
Race						
White	.97	(0.00)	.98	.97	(0.00)	.98
Smoking Behavior						
Current Smoker	.22	(0.00)	.19	.22	(0.00)	.16
Ever Smoker	.56	(0.00)	.54	.74	(0.00)	.71
N	4,648		1,574	3,556		1,242

Notes:
Columns 1 and 4 - Entire sample
Columns 3 and 6 - Part of sample with no item non-response and no attrition

Table 1.2: Health Measures – ELSA Wave 0

	Women			Men		
	1	2	3	4	5	6
Variable	Mean	(% missing)	Mean	Mean	(% missing)	Mean
Self-Rated Health						
Very bad/bad	.1	(0.00)	.06	.11	(0.00)	.06
Fair	.25	(0.00)	.21	.23	(0.00)	.19
Good	.39	(0.00)	.42	.37	(0.00)	.40
Very good	.27	(0.00)	.32	.29	(0.00)	.35
Medical Conditions						
Hypertension	.3	(0.11)	.23	.29	(0.03)	.24
Diabetes	.04	(0.02)	.02	.06	(0.00)	.04
Stroke	.03	(0.00)	.02	.04	(0.03)	.02
Heart Attack	.03	(0.02)	.01	.08	(0.03)	.04
Angina	.07	(0.00)	.04	.1	(0.03)	.05
Heart Murmur	.04	(0.02)	.04	.04	(0.00)	.02
Irregular Heart Rhythm	.07	(0.00)	.05	.08	(0.00)	.05
Other Heart Problems	.02	(0.02)	.01	.03	(0.00)	.01
GHQ12[1]	1.61	(6.26)	1.38	1.23	(6.21)	.92
No Longstanding Illness (LI)[2]	.45	(0.04)	.51	.43	(0.06)	.48
Non-limiting LI	.2	(0.04)	.21	.22	(0.06)	.24
Limiting LI	.36	(0.04)	.28	.36	(0.06)	.29
# Prescription drugs taken	1.85	(14.44)	1.31	1.62	(12.77)	1.1
Use of Medical Services						
# of GP visits last 2 weeks	.23	(0.02)	.2	.2	(0.03)	.16
# Hospital nights last year	1.03	(0.04)	.45	1.15	(0.08)	.51
Family Health History						
Father dead	.86	(2.32)	.83	.89	(2.53)	.88
Mother dead	.73	(1.61)	.67	.76	(1.77)	.71
At least one parent died of						
Hypertension	.01	(1.25)	.01	.01	(1.88)	.01
Angina	.03	(1.25)	.02	.02	(1.88)	.03
Heart Attack	.27	(1.25)	.27	.28	(1.88)	.27
Other Heart Problem	.14	(1.25)	.13	.15	(1.88)	.15
Stroke	.09	(1.25)	.09	.08	(1.88)	.09
Diabetes	.02	(1.25)	.03	.02	(1.88)	.02
Objective Health Data[3]						
Haemoglobin<13[a)], 11.5[b)] g/dL	.05	(32.66)	.03	.08	(27.33)	.06
Haemoglobin>18[a)], 16.5[b)] g/dL	.002	(32.66)	.001	.003	(27.33)	.001
Ferritin< 25[a)], 20[b)] μg/L	.14	(33.91)	.14	.06	(28.4)	.07
Ferritin>400[a)], 200[b)] μg/L	.03	(33.91)	.02	.02	(28.4)	.02
Total cholesterol>5 mmol/L	.8	(33.54)	.79	.75	(29.67)	.77
HDL cholesterol<1[a)], 1.2[b)] mmol/L	.14	(34.34)	.13	.18	(29.84)	.17
C-reactive protein>5 mg/L	.23	(32.22)	.21	.20	(27.18)	.16
Fibrinogen<1.7 g/L	.01	(41.61)	.01	.02	(37.46)	.02
Fibrinogen>3.7 g/L	.11	(41.61)	.11	.1	(37.46)	.08
Normal blood pressure untreated	.63	(22.93)	.75	.6	(22.41)	.72
Normal blood pressure treated	.17	(22.93)	.09	.18	(22.41)	.1
High blood pressure treated	.09	(22.93)	.05	.07	(22.41)	.04
Underweight (BMI<20)	.04	(11.79)	.03	.02	(10.46)	.01
Overweight (25≤BMI<30)	.38	(11.79)	.4	.52	(10.46)	.53
Obese (30≤BMI)	.25	(11.79)	.24	.21	(10.46)	.21
Waist-Hip-Ratio> 1[a)], 0.85[b)]	.29	(16.31)	.25	.15	(13.84)	.14
N	4,648		1,574	3,556		1,242

Notes:
Columns 1 and 4 - Entire sample, Columns 3 and 6 - Part of sample with no item non-response and no attrition. [1]12-item General Health Questionnaire, values range from 0 to 12. [2]For different longstanding illnesses see table 1.3. [3]Reference ranges taken from Oliveira (2008), [a)]Value for men, [b)]Value for women.

in wave 0 were asked "How is your health in general? Would you say it was very good, good, fair, bad, or very bad?". The columns 1 and 4 of table 1.2 display mean responses for the different SRH categories for women and men separately. As there are only few individuals who rate their health as bad or very bad, I group these two categories into one.

Tables 1.2 and 1.3 display the available health information. As column 2 for women and column 5 for men of table 1.2 show, self-reported health data and information on co-morbidities are broadly available. Objectively measured health data, however, is missing for a relatively large share of individuals. This results mainly from the fact that the measurement of objective health data is not compulsory. Instead, individuals can refuse to participate in the nurse visit, and even if they agree to the nurse visit they can refuse to have a blood sample taken. Overall, all health measures in wave 0 are available for about 50% of the sample.

When analyzing a new diagnosis or recurrence of a major health condition in the future the sample size shrinks further due to attrition. Information on the future diagnosis or recurrence of the health events is derived from an individual's answers to questions in the 2002, 2004 and 2006 waves of ELSA. The information is thus only available for individuals who appear again in ELSA after 1998. Overall, of the 4,305 individuals for whom the objective health data is available in 1998 only 71 percent are observed at least once in 2002, 2004 or 2006. For the analyses that focus on the future diagnosis or recurrence of major health conditions, there are thus two selection mechanisms that reduce the sample size. On the one hand, individuals have to participate in the nurse visit and have to have a blood sample taken. I call this selection mechanism 'no item non-response'. On the other hand, they have to stay in ELSA in the later waves, i.e. there is 'no attrition'.

When the analysis focuses on death, only the first of the two selection mechanisms is important. Information on death is collected regardless of attrition by linking the the data to information from the Department of Work and Pensions and to information contained in the National Health Service Central Register held by the Office of National Statistics.

As can be inferred from the differences in the means between columns 1 and 3 for women and between columns 4 and 6 for men in tables 1.1 and 1.2 the individuals who remain in the sample when item non-response in wave 0 and attrition are taken into account differ in many aspects from the overall sample. They are on average younger, more likely to be white or married and less likely to be smokers. Also, they are healthier in terms of both subjective and objective health measures. If the predictive

Table 1.3: Prevalence of Different Longstanding Illnesses – Wave 0

	Women		**Men**	
	1	2	3	4
Variable	Mean	Mean	Mean	Mean
Cancer	.03	.02	.02	.01
Endocrine/metabolic disorders	.05	.05	.02	.02
Mental illness	.03	.02	.02	.02
Migraine/headaches	.02	.02	.01	.01
Other problem nervous system	.03	.02	.03	.02
Cataract/poor eye sight	.02	.01	.02	.02
Other eye problems	.01	.01	.02	.02
Poor hearing/deafness	.02	.01	.03	.03
Other ear complaints	.04	.01	.05	.02
Complaints of blood vessels	.01	.01	.02	.01
Bronchitis/emphysema	.01	.01	.02	.01
Asthma	.06	.06	.05	.06
Respiratory complaints	.02	.02	.03	.03
Stomach ulcer	.03	.02	.03	.04
Other digestive complaints	.02	.02	.01	.01
Complaints of bowel/colon	.04	.03	.02	.02
Reproductive system disorders	.01	.02	.02	.02
Arthritis	.18	.15	.12	.12
Back problems	.06	.06	.08	.08
Problems of bones/joints/muscles	.07	.06	.08	.09
Skin complaints	.01	.01	.02	.01
N	4,648	1,574	3,556	1,242

Notes: Information on specific longstanding illness is missing for 0.09% of women and 0.06% of men.
Columns 1 and 3 - Entire sample
Columns 2 and 4 - Part of sample with no item non-response and no attrition

power of SRH for subsequent health events varies with health or age, ignoring the two selection mechanisms might result in biased estimates. Furthermore, the existence of unobservable influences on selection that also affect the future health outcomes would lead to inconsistent estimates when ignoring selection. Inverse probability weighting is employed to correct for the two selection mechanisms (see appendix to this chapter for details).

Figure 1.1: SRH and Future Health Events - Men

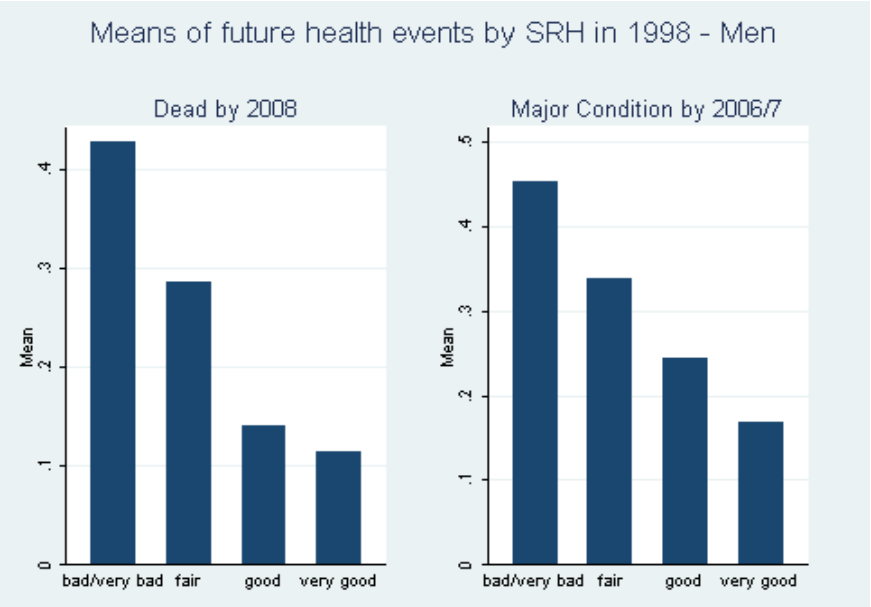
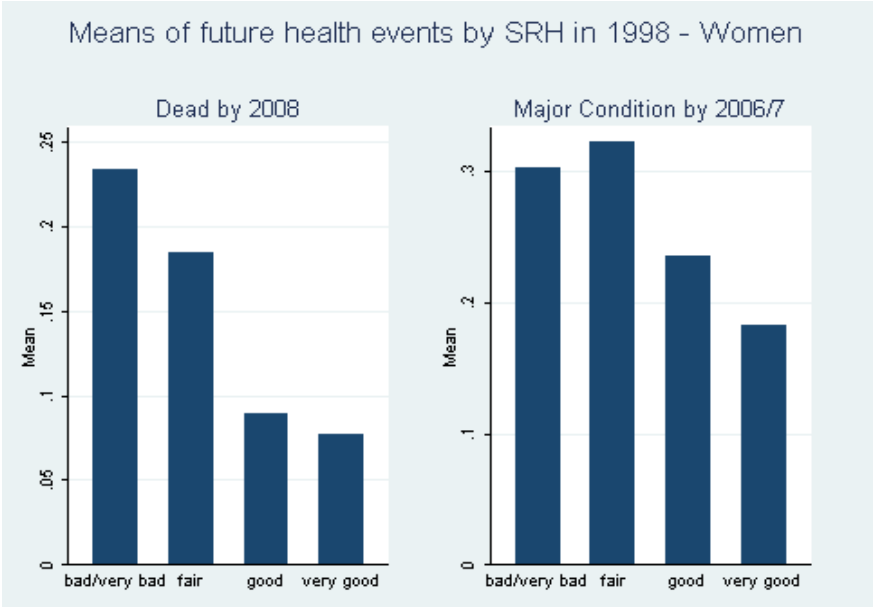


Figure 1.2: SRH and Future Health Events - Women



Figures 1.1 and 1.2 present a first glance at the relationship between SRH and the future health events. There is a graded relationship of SRH and all-cause mortality for both genders: The better SRH in 1998, the lower the proportion of individuals who are dead by the year 2008. For men, a similar graded relationship can be observed between SRH and a diagnosis of a major condition within the next 8 years. For women, the graded relationship holds for the categories of fair, good, and very good SRH.

1.4 Estimation Strategy

The existence of private information in health insurance, life insurance and annuity markets is investigated by regressing indicator variables for the occurrence of future health events on categories of SRH and different control variables in a baseline year. I include different sets of control variables to capture the information used for underwriting by insurance companies in different insurance markets. The future health events are meant to capture the future realization of risks. The realization of the risk that is insured in life insurance and annuity markets is captured by an indicator for whether an individual is dead 10 years after the baseline interview. The realization of the insured risk in health insurance markets is captured by a variable that indicates whether an individual is newly diagnosed or has a recurrence of heart disease, cancer or stroke within 8 years after the initial interview.

The information on the future health events that is contained in SRH is interpreted as evidence for private information. SRH might, of course, not capture all private knowledge on health and mortality risk. SRH asks about health at the time of the interview and not about expected death or changes in health in the future. A better suited proxy for private information might be subjective life expectancy.[5] This variable with this information, however, is only elicited starting from ELSA wave 1. As objectively measured health information is only available in wave 0 and 2, using subjective life expectancy in the analysis would considerably reduce follow-up time. Subjective life expectancy is, however, used as a proxy for individual information as a robustness check.

Let y_i^{j*}, $j \in \{D, M\}$ denote latent variables for the future health events, where D stands for death and M for being diagnosed with a major condition, for individual i.

[5]Hurd and McGarry (2002) find that subjective life expectancy contains an expectational component in addition to the health component that is also captured by SRH.

Each of the y_i^{j*}'s can be represented by the following equations

$$\begin{aligned} y_i^{j*} &= \beta_0^j + \beta_1^j SRH_{1i} + \beta_2^j SRH_{2i} + \beta_3^j SRH_{3i} + \beta_4^j X_{ai} + \epsilon_i^j \\ y_i^j &= \mathbb{I}(y_i^{j*} > 0) \end{aligned} \tag{1.1}$$

where SRH_k, $k \in \{1, 2, 3\}$, represent dummies for the three categories, very bad/bad, good and very good SRH with fair SRH as the reference category. X_{ai} is a vector of variables that represents the information that insurance companies in insurance market a collect and use for underwriting.

ϵ_i^j captures unobservables influences on the latent future health event j. Under the assumptions that $\epsilon_i^j \sim N(0, 1)$ and that the correlation between ϵ^D and ϵ^M, ρ_{DM}, is equal to 0, I estimate $Pr(y_i^j = 1)$ for each of the health events independently using single equation probit models.[6]

Table 1.4 displays the variables that are typically collected and used in the application process in different insurance markets in the UK and in the US. The listed variables represent information that is used for risk classification and calculation of premia. As Finkelstein and Poterba (2006) observed, insurance companies may have additional information about applicants that is not used in underwriting. These "unused observables" do not mitigate the scope for adverse selection and are thus not included as controls in X.

The vector of control variables for the investigation of private information on the insured risk in group health insurance markets, $X_{GroupHI}$, does not include any variables. In the US employer-sponsored health insurance, federal law forbids individual underwriting (GAO, 2003). In the UK, employer-sponsored private health insurance also refrains from individual underwriting (Mossialos and Thomson, 2009). In annuity markets, underwriting is based on the individual's age and sex. As I estimate separate models for men and women the vector of control variables in annuity markets, $X_{Annuities}$, includes only information on age.

Considerably more information is used for underwriting in life and individual health insurance. In addition to age and sex, information on lifestyle is typically used. Age and information on an individual's smoking history are thus included in the vector of control variables. Furthermore, medical information of individuals and sometimes also their families can be used for underwriting. The insurers in the UK and some

[6]The assumption of independence between the error terms can be relaxed and a bivariate equation probit models be estimated. For this analysis, however, the assumption of independent error terms seems appropriate as only private information on the particular risk insured in each specific market, i.e. either the risk of dying/surviving or the risk of high medical expenses, is of interest.

Table 1.4: Information Used in Underwriting in US and UK Insurance Markets

	Health Risk		Mortality Risk	
	Group HI[a]	Individual HI[b]	Annuities[c]	Life Insurance[d]
Age		✓	✓	✓
Sex		✓	✓	✓
Smoking Behavior		✓		✓
Medical Conditions		✓		✓
Objective Health Data		✓		✓
Use of Prescription Drugs		✓		✓
Prior Use of Medical Services		✓		
Family Health History		(✓)		✓
Occupational Class		✓		✓
Alcohol/Substance Abuse		✓		✓
Driving Information		✓		✓
Residence/Citizenship		✓		✓
Dangerous Hobbies		✓		✓
Foreign Travel		✓		✓

Notes:
a) See GAO (2003) and "Risk Classification" (1999). *b*) See "Risk Classification" (1999). *c*) See Cutler et al. (2008a). *d*) See Cutler et al. (2008a) and He (2009). The table displays types of information used in risk classification and calculation of premia in different insurance markets. While in the US family history is typically not used in underwriting health insurance ("Risk Classification" 1999), it is used in the UK .

US states can even require insurance applicants to undergo medical examinations. They thus have clinical information on blood values and other objectively measured information. To capture the medical information I include self-reported conditions, linear splines of objectively measured BMI, waist-to-hip ratio, and blood values in X_{Life} and $X_{IndividualHI}$. I further use parents' cause of death as a proxy for familial medical history.

Insurance companies sometimes use information on occupational status, dangerous occupations, hazardous hobbies, risky travel destinations, residence/citizenship, and alcohol or drug abuse for underwriting. As a proxy for occupation, I include dummies for occupational social class as represented in table 1.1. The other conditions are unfortunately not well captured in the ELSA data.

1.5 Results

Average marginal effects of the SRH categories after estimating equation (1.1) are reported in tables 1.5 to 1.8. The first two tables display average marginal effects of the SRH categories on whether an individual dies within the next 10 years for men and women, respectively. Tables 1.7 and 1.8 display average marginal effects of SRH on a new diagnosis or recurrence of one of the major health conditions, heart disease, cancer or stroke, in the next 8 years. All tables show results that are weighted to correct for missing data.[7]

Column 2 in each of the tables 1.5 and 1.6 is of specific interest. In addition to SRH the estimated models include the annuity underwriting controls. For both genders, all three SRH coefficients are significantly different from 0. Conditional on age, women in very good SRH in 1998 are 8 percentage points less likely to be dead in 2008 than women who rate their health as fair. Similarly, men in very good SRH in 1998 are 13 percentage points less likely to be dead in 2008 than men in fair SRH. When interpreting conditional information in SRH as private information these results show evidence of a scope for adverse selection in annuity markets.

From left to right in tables 1.5 and 1.6 more control variables are added. The last columns report average marginal effects of the SRH categories when age, smoking information, medical conditions, family health history, social occupational class, and objective health data are included as controls. The p-values of the Wald test for joint significance indicate that for both genders the three SRH coefficients are still jointly significantly different from 0 at the 10 percent significance level. The inclusion of the additional controls, however, and in particular the inclusion of the objectively measured health information, leads to an attenuation in the average marginal effects and to a reduction in the significance of the underlying SRH coefficients. Thorough underwriting, and in particular medical underwriting that includes blood tests and other objectively measured health data, is thus able to considerably reduce private information about mortality risks.

Tables 1.7 and 1.8 show similar results for the diagnosis or re-diagnosis of a major health condition within 8 years after the baseline interview. The results presented in column 1 of each table indicate that with no controls added in addition to SRH, the coefficient of the three SRH categories are jointly significantly different from zero for women and men. When there is no individual underwriting, as in the case of group

[7]The estimation of the weights is described in the appendix to this chapter. Unweighted results do not differ significantly from the weighted results.

Table 1.5: Private Information on Mortality Risk – Women

Controls	1	2	3	4	5	6	7
SRH – very bad/bad	.144**	.156***	.141***	.066*	.074**	.075**	.056*
	(.075)	(.055)	(.049)	(.04)	(.04)	(.04)	(.035)
SRH – good	-.14***	-.06***	-.054***	-.04**	-.041**	-.041**	-.029
	(.031)	(.022)	(.021)	(.021)	(.021)	(.021)	(.019)
SRH – very good	-.15***	-.068***	-.061***	-.042*	-.043*	-.045*	-.029
	(.031)	(.023)	(.022)	(.023)	(.024)	(.024)	(.023)
Age splines		✓	✓	✓	✓	✓	✓
Smoking Behavior			✓	✓	✓	✓	✓
Medical Conditions				✓	✓	✓	✓
Family Health History					✓	✓	✓
Occupational Class						✓	✓
Objective Health Data							✓
Wald test (p-value)	.000	.000	.000	.019	.012	.009	.070
N	2,184	2,184	2,184	2,155	2,132	2,132	2,132
Pseudo R^2	0.0679	0.3030	0.3266	0.3712	0.3811	0.3877	0.4256

* p<0.10, ** p<0.05, *** p<0.01 - test of the underlying coefficient being 0

Notes: Dependent variable is 1 if individual is dead by 2008. Reference category SRH: fair. Average marginal effects after probit estimation calculated as $\frac{1}{N}\sum_i \left[\Phi(\widehat{\beta_0}+\widehat{\beta_k}+\widehat{\beta_4}X_i)-\Phi(\widehat{\beta_0}+\widehat{\beta_4}X_i)\right]$ for $k \in \{1,2,3\}$. Standard errors of the marginal effects calculated using the delta method in parentheses. Estimates are weighted to correct for selection due to missing objective health data. Wald test p-values for testing the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. Medical Conditions include number of prescription drugs taken. The number of observations changes between columns as observations for which some explanatory variables result in perfect prediction in the probit estimation are automatically dropped.

Table 1.6: Private Information on Mortality Risk – Men

Controls	1	2	3	4	5	6	7
SRH – very bad/bad	.274***	.228***	.212***	.104**	.102**	.094**	.064
	(.064)	(.053)	(.051)	(.05)	(.049)	(.048)	(.045)
SRH – good	-.164***	-.096***	-.088***	-.062**	-.057**	-.054**	-.05**
	(.034)	(.027)	(.026)	(.026)	(.026)	(.026)	(.025)
SRH – very good	-.203***	-.132***	-.119***	-.06**	-.054*	-.05*	-.035
	(.033)	(.027)	(.026)	(.029)	(.029)	(.028)	(.028)
Age splines		✓	✓	✓	✓	✓	✓
Smoking Behavior			✓	✓	✓	✓	✓
Medical Conditions				✓	✓	✓	✓
Family Health History					✓	✓	✓
Occupational Class						✓	✓
Objective Health Data							✓
Wald test (p-value)	.000	.000	.000	.003	.006	.011	.044
N	1,766	1,766	1,766	1,766	1,766	1,766	1,766
Pseudo R^2	0.1051	0.3206	0.3350	0.3961	0.4043	0.4088	0.4389

* p<0.10, ** p<0.05, *** p<0.01 - test of the underlying coefficient being 0

Notes: Dependent variable is 1 if individual is dead by 2008. Reference category SRH: fair. Average marginal effects after probit estimation calculated as $\frac{1}{N}\sum_i \left[\Phi(\widehat{\beta_0}+\widehat{\beta_k}+\widehat{\beta_4}X_i)-\Phi(\widehat{\beta_0}+\widehat{\beta_4}X_i)\right]$ for $k \in \{1,2,3\}$. Standard errors of marginal effects calculated using the delta method in parentheses. Estimates are weighted to correct for selection due to missing objective health data. Wald test p-values for testing the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. Medical Conditions include number of prescription drugs taken.

Table 1.7: Private Information on Health Risk – Women

Controls	1	2	3	4	5	6	7	8
SRH – very bad/bad	.083	.137	.133	-.084	-.088	-.085	-.088	-.093*
	(.144)	(.103)	(.1)	(.049)	(.05)	(.055)	(.049)	(.046)
SRH – good	-.182***	-.091**	-.091**	.011	.011	.01	.009	.011
	(.052)	(.042)	(.042)	(.035)	(.035)	(.034)	(.035)	(.033)
SRH – very good	-.256***	-.167***	-.168***	-.024	-.023	-.02	-.022	-.011
	(.052)	(.043)	(.043)	(.039)	(.039)	(.039)	(.039)	(.038)
Age splines		✓	✓	✓	✓	✓	✓	✓
Smoking Behavior			✓	✓	✓	✓	✓	✓
Medical Conditions				✓	✓	✓	✓	✓
Service Use					✓	✓	✓	✓
Family Health History						✓	✓	✓
Occupational Class							✓	✓
Objective Health Data								✓
Wald test (p-value)	.000	.000	.000	.248	.238	.284	.273	.244
N	1,572	1,572	1,572	1,572	1,572	1,572	1,572	1,571
Pseudo R^2	0.0519	0.1381	0.1389	0.2551	0.2557	0.2636	0.2651	0.2801

* p<0.10, ** p<0.05, *** p<0.01 - test of the underlying coefficient being 0

Notes: Dependent variable is 1 if individual diagnosed or re-diagnosed with heart disease, cancer or stroke in waves 1,2 or 3. Reference category SRH: fair. Average marginal effects after probit estimation calculated as $\frac{1}{N}\sum_i \left[\Phi(\widehat{\beta_0}+\widehat{\beta_k}+\widehat{\beta_4}X_i)-\Phi(\widehat{\beta_0}+\widehat{\beta_4}X_i)\right]$ for $k \in \{1,2,3\}$. Standard errors of marginal effects calculated with the delta method in parentheses. Estimates are weighted to correct for selection due to missing objective health data and attrition. Wald test p-values for testing the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. Medical Conditions include number of prescription drugs taken. The number of observations changes between columns as observations for which some explanatory variables result in perfect prediction in the probit estimation are automatically dropped

Table 1.8: Private Information on Health Risk – Men

Controls	1	2	3	4	5	6	7	8
SRH – very bad/bad	.106	.068	.051	-.055	-.047	-.028	-.031	-.025
	(.119)	(.108)	(.104)	(.064)	(.064)	(.062)	(.06)	(.056)
SRH – good	-.068	-.051	-.051	.009	.009	.012	.012	.023
	(.063)	(.062)	(.059)	(.043)	(.043)	(.041)	(.041)	(.038)
SRH – very good	-.119*	-.107*	-.112*	-.026	-.024	-.008	-.012	-.003
	(.067)	(.067)	(.065)	(.047)	(.047)	(.045)	(.045)	(.041)
Age splines		✓	✓	✓	✓	✓	✓	✓
Smoking Behavior			✓	✓	✓	✓	✓	✓
Medical Conditions				✓	✓	✓	✓	✓
Service Use					✓	✓	✓	✓
Family Health History						✓	✓	✓
Occupational Class							✓	✓
Objective Health Data								✓
Wald test (p-value)	.093	.170	.180	.652	.695	.893	.849	.771
N	1,240	1,240	1,240	1,240	1,240	1,240	1,240	1,239
Pseudo R^2	0.0224	0.0650	0.0784	0.2217	0.2243	0.2495	0.2520	0.2746

* p<0.10, ** p<0.05, *** p<0.01 - test of the underlying coefficient being 0

Notes: Dependent variable is 1 if individual diagnosed or re-diagnosed with heart disease, cancer or stroke in waves 1,2 or 3. Reference category SRH: fair. Average marginal effects after probit estimation calculated as $\frac{1}{N}\sum_i \left[\Phi(\widehat{\beta_0}+\widehat{\beta_k}+\widehat{\beta_4}X_i)-\Phi(\widehat{\beta_0}+\widehat{\beta_4}X_i)\right]$ for $k \in \{1,2,3\}$. Standard errors of marginal effects calculated with delta method in parentheses. Estimates are weighted to correct for selection due to missing objective health data and attrition. Wald test p-values for testing the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. Medical Conditions include number of prescription drugs taken. The number of observations changes between columns as observations for which some explanatory variables result in perfect prediction in the probit estimation are automatically dropped.

health insurance, there is thus private information and therefore scope for adverse selection.

The inclusion of additional controls from left to right in tables 1.7 and 1.8 results in a reduction in the information in SRH. For both genders, the inclusion of self-reported medical conditions and the number of prescription drugs taken as shown in columns 4 of the respective tables leaves no additional explanatory power in SRH for the diagnosis or re-diagnosis of a major health condition. Taking information in SRH on a future diagnosis of a major condition as private information, the results indicate that medical underwriting is a crucial determinant of the degree of private information left in private health insurance markets.

Overall, the results present evidence that medical underwriting, and in particular underwriting including medical examinations significantly reduces private information as captured by SRH. While I find evidence that private information on mortality risks remains even with stringent underwriting, I find no evidence of private information on health risks when stringent underwriting is employed. There is evidence, however, for private information on health risks when no or scarce information is used in underwriting. Differences in adverse selection between group health insurance markets without individual underwriting and individual health insurance with stringent individual underwriting can therefore be reconciled precisely by the differences in underwriting.

1.6 Robustness Analyses

In this section, I present different sensitivity analyses. First, the time horizon within which the realization of risk can occur is shortened to 4 years to investigate whether individuals have more information on risks in a shorter term. Second, subjective life expectancy is included in the analysis instead of SRH, as it might be a better proxy for private information on future health events. Third, I analyze how well the underwriting controls help to predict the outcomes to gauge the scope for private information independent of the different proxies for individual information.

Table 1.9 reports the average marginal effects of the three SRH categories when the dependent variables capture the realization of risk within the next 4 years. Results for women are displayed in the upper panel, results for men in the lower panel. The dependent variable in columns 1 and 2 is an indicator for whether an individual is dead by ELSA wave 1. In columns 3 and 4 the dependent variable is an indicator for whether an individual reports a diagnosis or recurrence of a major health condition in wave 1. The first column for each event includes a limited set of control variables in addition to

Table 1.9: Robustness – Event by 2002

	Mortality Risk		**Health Risk**	
	Annuities	Life Insurance	Group HI	Individual HI
	1	2	3	4
Women				
SRH – very bad/bad	.094***	.036	.114	-.032
	(.037)	(.025)	(.163)	(.043)
SRH – good	-.022	-.002	-.176***	-.003
	(.016)	(.014)	(.055)	(.029)
SRH – very good	-.022	.002	-.22***	-.004
	(.016)	(.016)	(.055)	(.033)
Wald test (p-value)	0.000	0.407	0.000	0.917
N	2,184	2,054	1,558	1,557
Pseudo R^2	0.2398	0.4303	0.0660	0.3427
Men				
SRH – very bad/bad	.103**	-.004	.136	.005
	(.052)	(.031)	(.116)	(.051)
SRH – good	-.036*	-.015	-.079	-.014
	(.02)	(.02)	(.055)	(.033)
SRH – very good	-.069***	-.031	-.139**	-.076**
	(.02)	(.022)	(.057)	(.034)
Wald test (p-value)	0.000	0.582	0.008	0.08
N	1,766	1,690	1,222	1,204
Pseudo R^2	0.2094	0.3719	0.0507	0.3442

* p<0.10, ** p<0.05, *** p<0.01 - test of the underlying coefficient being 0

Notes: Average marginal effects after probit estimation and their standard errors in parentheses. Different columns include different sets of control variables. Column Annuities includes age splines, Life Insurance includes age splines, smoking information, medical conditions, number of prescription drugs taken, family health history, social occupational class, and objective health data, Group HI includes no controls in addition to SRH, and Individual HI includes the same variables as Life Insurance plus use of medical services. Results are weighted to correct for missing data.

SRH, the second column includes the most comprehensive set of underwriting controls.

Similarly to the results of the main analysis, there is significant information in SRH on dying within the next 4 years and on being diagnosed with a major health condition within the next 4 years when only limited sets of underwriting controls are included. Including the most comprehensive set of the life insurance underwriting controls, X_{Life}, eliminates any information in SRH on the 4 year mortality risk. With respect to the 4 year health risk, however, column 4 of table 1.9 presents evidence that at least for men some information remains in SRH even when the set of individual health insurance controls, $X_{IndividualHI}$, is included. SRH thus may contain more information in the shorter term.

In the upper panel of table 1.10 results are reported with subjective life expectancy replacing SRH in equation (1.1) as proxy for private information.[8] Subjective life expectancy is only available in ELSA starting from wave 1. The earliest wave that contains subjective life expectancy and objectively measured health data is wave 2. I thus use wave 2 as baseline in this robustness analysis. The dependent variables are whether an individual is dead by 2008 and whether an individual reports a diagnosis or recurrence of heart disease, cancer or stroke in wave 3. As wave 2 was elicited in 2006, the time horizon within which the risks can materialize is 2 years.

For means of comparison results using SRH in the 2006 wave of ELSA as proxy for individual information are reported in the lower panel of table 1.10. Columns 1 and 2 report average marginal effects on the probability of being dead in 2008 for men and women, respectively, columns 3 and 4 report average marginal effects on the probability of being diagnosed with heart disease, cancer, or stroke for the two genders. Only results with the most comprehensive sets of additional control variables are presented in order to gauge how much private information there is in the case of stringent underwriting.

The results presented in the upper panel of table 1.10 indicate that subjective life expectancy contains information on whether individuals die or survive the next 2 years even when a comprehensive set of additional controls is included. Column 3 and 4 present evidence that there is no significant information in subjective life expectancy for an onset of a major health condition within the next 2 years. SRH in 2006, on the contrary, contains information on dying in the next two years only for men and on the onset of a major health condition for both genders. These results suggest that subjective life expectancy contains more information on mortality risks, while SRH is

[8]Subjective life expectancy in ELSA is elicited with the question "What are the chances that you will live to be x or more?" Where x depends on the individual's age at the time of the interview. The average time horizon for this question in the estimation sample is 15 years.

Table 1.10: Robustness – Subjective Life Expectancy

	Mortality Risk		**Health Risk**	
	Men	Women	Men	Women
	1	2	3	4
Subjective Life Expectancy				
Probability of surviving – $0-24\%$	-.032	.004	.049	.028
	(.020)	(.014)	(.037)	(.036)
Probability of surviving – $50-74\%$	-.056***	-.023**	-.019	.008
	(.019)	(.012)	(.030)	(.029)
Probability of surviving – $75-100\%$	-.054***	-.017	-.016	-.025
	(.020)	(.014)	(.031)	(.030)
Wald test (p-value)	0.003	0.0156	0.0968	0.1358
N	2,136	2,504	1,909	2,315
Pseudo R^2	0.3073	0.2814	0.2472	0.2838
Self-Rated Health				
SRH – bad/fair	.046***	.010	.027	.071***
	(.016)	(.009)	(.028)	(.024)
SRH – very good	-.011	-.012	-.026	.007
	(.010)	(.007)	(.022)	(.019)
SRH – excellent	.005	-.007	-.060**	-.015
	(.016)	(.011)	(.028)	(.025)
Wald test (p-value)	0.004	0.1312	0.0920	0.012
N	2,135	2,504	1,908	2,315
Pseudo R^2	0.3081	0.2734	0.2464	0.2866

* p<0.10, ** p<0.05, *** p<0.01 - test of the underlying coefficient being 0

Notes: Average marginal effects after probit. Standard errors in parentheses. Dependent variable whether an individual is dead by the year 2008 for mortality risk, and whether there was an onset or recurrence of heart disease, cancer and/or stroke for health risk. Explanatory variables are taken from ELSA wave 2. Reference category of subjective life expectancy is 25-49%. SRH measured on the US scale in wave 2, good SRH, the middle category, chosen as reference. Control variables in each estimation are: age splines, smoking information, medical conditions, number of prescription drugs taken, family health history, and objective health data. Results are weighted to correct for missing objective health data using weights provided in ELSA.

better suited to predict health risks – at least in the short run. Whether this finding also holds in the longer term, is subject to additional analyses that will be possible once a longer follow-up period is available in ELSA.

SRH and subjective life expectancy are only proxies for individual information on future health risks. Individuals could have information on the insured risks that is neither included in SRH nor subjective life expectancy nor in the verifiable underwriting controls. In order to gauge the scope for private information independent of a proxy for individual information, I estimate probit models where only the underwriting controls are included

$$\begin{aligned} y_i^{j*} &= \gamma_0^j + \gamma_1^j X_{ai} + \eta_i^j \\ y_i^j &= \mathbb{I}(y_i^{j*} > 0). \end{aligned} \tag{1.2}$$

X_{ai} represents the vectors of the respective most comprehensive sets of underwriting controls for both dependent variables, y_i^M and y_i^D. The better these comprehensive sets of underwriting controls help to predict the future realization of the insured risk, the better the insurer will be able to discriminate between risk types based on the collected information. The less scope thus remains for private information.

In order to analyze how well models explain the variation in binary dependent variables, often the percent of correctly predicted observations is reported. This approach involves predicting the probability, $\widehat{Pr}(y_i^j = 1|X_{ai})$, for each individual and choosing some cutoff above which the researcher assumes that the predicted probability corresponds to predicting a 1 and below which it corresponds to predicting a 0. One can then calculate how many of the actual ones and how many of the actual zeros are correctly predicted by the model. It is not clear, however, what the optimal choice of cutoff is (see Wooldridge, 2010, p. 573f. for a discussion).

In order to avoid the choice of a specific cutoff, I display the distributions of the predicted probabilities by the values of the actual outcome variable. For example, figure 1.3 contains the cumulative distribution functions of $\widehat{Pr}(y_i^D = 1|X_{ai})$ for men who are still alive in 2008 and for men who are dead by 2008. Figure 1.4 displays the same for women, and figures 1.5 and 1.6 display the respective cumulative distribution functions of $\widehat{Pr}(y_i^M = 1|X_{ai})$ for men and women, respectively.

For any cutoff probability, the figures show how large the share is of individuals in the two outcome groups with a predicted probability below or above the respective cutoff. In figure 1.3 with a cutoff at $\widehat{Pr}(y_i^D = 1|X_{ai}) = 20\%$, for example, 80% of men who are still alive in 2008 have a predicted probability that is lower than the cutoff,

Figure 1.3: Predicted Probability Death - Men

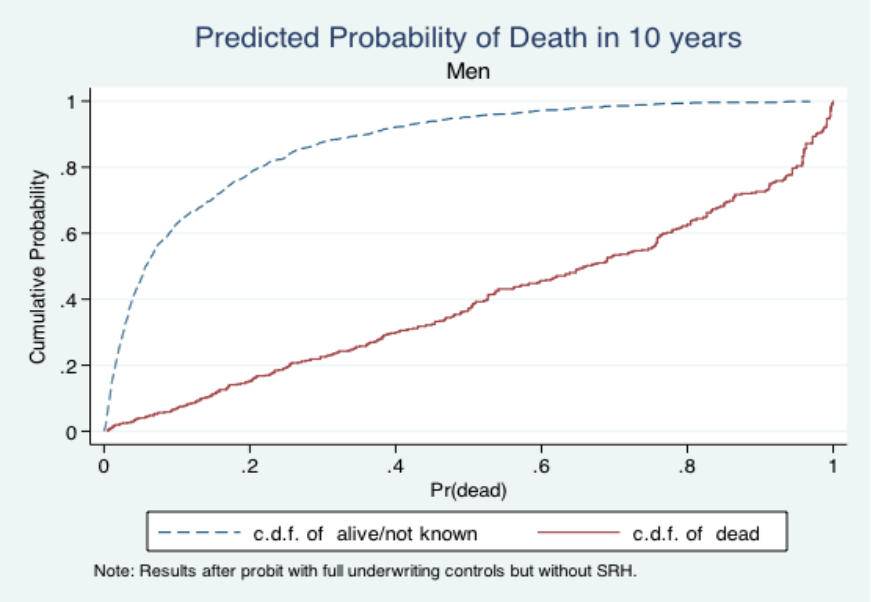
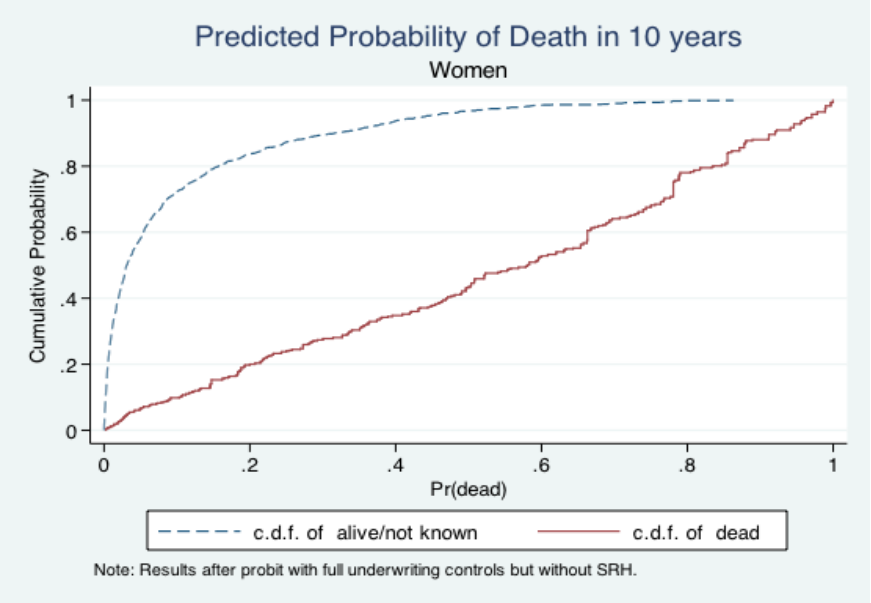


Figure 1.4: Predicted Probability Death - Women



and 80% of men who are dead by 2008 have a predicted probability that is higher than the cutoff. If an insurer thus set a cutoff at 20% and treated every individual with a predicted probability below this cutoff as low risk and every individual with a predicted probability above this cutoff as high risk, it would correctly classify 80% of men. Figures 1.4 to 1.6 display similar patterns. Based on the underwriting controls it thus seems to be possible to discriminate rather accurately between individuals who experience an event in the future and individuals who do not experience the respective event. This

Figure 1.5: Predicted Probability Major Condition - Men

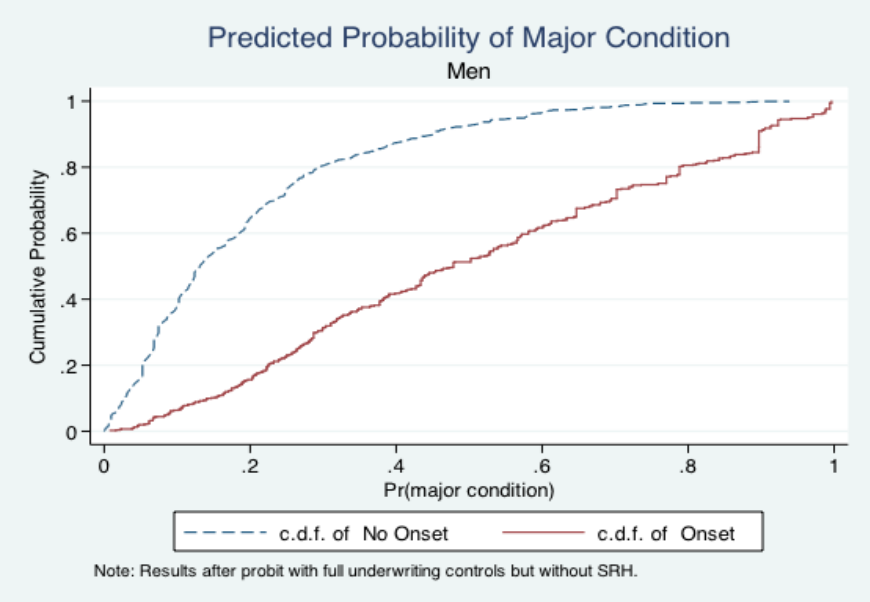
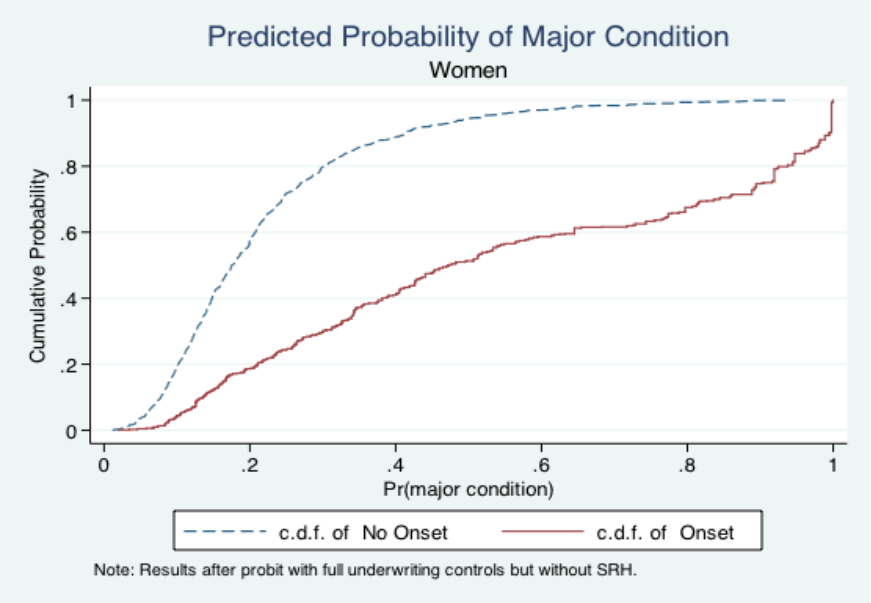


Figure 1.6: Predicted Probability Major Condition - Women



provides further evidence that there is not a lot of room for private information in the different insurance markets when stringent underwriting is employed.

Overall, the analyses presented in this section strengthen the intuition that stringent underwriting helps to reduce private information in the markets for life insurance and individual health insurance. At least for the individuals' information on the risks insured in health insurance markets, SRH seems to be a better proxy than subjective life expectancy. Furthermore, there is evidence that even independent of the choice

of a proxy for individuals' information on their health risks, little scope for private information remains with stringent underwriting.

1.7 Conclusion

Mixed empirical evidence on adverse selection in different insurance markets stands in contrast to robust predictions of the existence of adverse selection in theoretical models. This divergence has engendered research that tries to reconcile theoretical and empirical results. This chapter focuses on one possible explanation for the mixed findings in the empirical literature: different scopes of adverse selection in different insurance markets. In insurance markets in which insurance companies collect a lot of information about their applicants and use this information for underwriting, private information might not exist and thus there might be no scope for adverse selection.

In this study, I focus on the markets for life insurance, annuities and health insurance. Private information in these markets is detected by using information in SRH on dying in the next years and on future diagnoses of diseases. Different verifiable measures are included as controls in addition to SRH to imitate different types of information used in underwriting in the different insurance markets. The information that remains in SRH when the different sets of underwriting controls are included is interpreted as private information in the different insurance markets.

The analysis employs data from ELSA, one of the rare longitudinal data sets that provides objectively measured health data. The data set allows to mimic medical underwriting with greater precision than is typically possible with data from other population surveys that only include self-reported health information. As thorough underwriting that includes outcomes of medical examinations is typically conducted in life insurance markets and private individual health insurance markets, ELSA's objectively measured health data is particularly valuable for my analysis.

Dying within the next 10 years is significantly related to SRH at baseline when only information on age and sex is additionally included in the analysis. When medical information and in particular objectively measured health data are additionally included, however, the predictive power of SRH for death or survival is considerably reduced. Similarly, there is significant information in SRH for the diagnosis or recurrence of a major health condition in the next 8 years with only limited controls. The inclusion of medical information reduces the amount of information contained in SRH to insignificant levels.

Interpreting the remaining information in SRH when underwriting controls are

included in the analyses as private information, I thus find evidence for private information on mortality risk and health risk if insurance companies do not use medical information for underwriting. When medical underwriting is conducted, however, the private information and thus the scope for adverse selection is reduced in life insurance and eliminated in health insurance markets.

These findings can help to reconcile differences in the evidence of adverse selection between life insurance and annuity markets and between group and individual private health insurance markets. Only limited information is used for underwriting in group health insurance and annuity markets and these markets are found to suffer from adverse selection. In individual health and life insurance markets medical underwriting is typically performed and there is little evidence for adverse selection in these markets. Differences in the information used for underwriting in the different markets – partly resulting from different underwriting regulations – leave different amounts of private information and thus different scopes for adverse selection.

My results in turn indicate that employing stringent underwriting in group health insurance and annuity markets could reduce and possibly eliminate adverse selection in these markets. Recent research finds that adverse selection reduces welfare in the market for annuities (Einav et al., 2010b) and in health insurance markets (Bundorf et al., 2008). A question that my analysis cannot answer, however, is how stringent underwriting would affect overall welfare in these markets. On the one hand, stringent underwriting might increase welfare by eliminating welfare losses resulting from adverse selection. On the other hand, it might at the same time reduce welfare in other dimensions. For example, in the case of short-term insurance contracts where individuals have to buy insurance coverage repeatedly, repeated stringent underwriting would introduce an additional risk to the existing health-related risks, namely the risk of becoming a 'high risk' in the future. If there is no insurance to insure against this additional risk, a "welfare loss from incomplete insurance contracts" will arise (Cutler and Zeckhauser, 2000, p. 627). Whether stringent underwriting increases or decreases welfare thus depends on the specific institutional setting.

Appendix: Inverse Probability Weighting

In order to correct for item non-response and attrition I use inverse probability weighting. To correct for the two selection mechanisms simultaneously, I estimate the joint probability of availability of all objective health measures in wave 0, $(h = 1)$, and no attrition after wave 0, $(a = 1)$, using a bivariate probit model.

The crucial assumption for consistency with this approach is conditional independence

$$Pr(a = 1, h = 1|y, SRH, X_{IndividualHI}, d) = Pr(a = 1, h = 1|SRH, Z, d)$$

where $Z \subset X_{IndividualHI}$, and $X_{IndividualHI}$ is the most comprehensive set of underwriting controls. Z includes all variables in $X_{IndividualHI}$ except for the objectively measured health data. d is a vector of additionally included control variables.

The conditional independence assumption would be invalidated by the existence of unobservables that influence both – selection and the outcome y. The variables included in d thus not only have to be significant predictors of attrition and item non-response but also have to be related to y in order to attenuate the worry of unobservable influences. Potential candidates for inclusion in d are health-related variables that are not used by insurance companies for underwriting and are therefore not included in the different sets of control variables X.

In my analysis, d includes information on marital status, activity status, race, the household's economic situation, survey participation behavior of other survey members in the same household, the individual's survey participation behavior in other parts of the survey in wave 0, and information on the situation during the interview.

Separate bivariate probit models are estimated for men and women. The coefficients are displayed in table 1.11. Many of the variables included in d significantly affect selection. Furthermore, the two selection mechanisms are positively correlated for both genders. An individual who is more likely to stay in the survey is also more likely to have a nurse visit and a blood sample taken. This could reflect an unobserved liking for surveys.

The results in table 1.11 are used to predict $Pr(a = 1, h = 1|SRH, Z, d)$ and $Pr(h = 1|SRH, Z, d)$ for each individual. $\frac{1}{\hat{Pr}(a=1,h=1|SRH,Z,d)}$ is used as a weight in estimation when the dependent variables is an indicator of a diagnosis or reccurence of a major health condition. Whether an individual dies or not is observed irrespective of attrition, therefore $\frac{1}{\hat{Pr}(h=1|SRH,Z,d)}$ is used as a weight in the estimations for private information on mortality risk.

Table 1.11: Selection Mechanisms – Bivariate Probit

	Women				Men			
	a		h		a		h	
Completed other	0.197	(0.167)	0.377**	(0.157)	-0.146	(0.202)	0.465***	(0.180)
Partner a=1	2.360***	(0.067)	0.262***	(0.049)	2.468***	(0.071)	0.235***	(0.053)
Partner h=1	-0.154**	(0.064)	0.638***	(0.049)	-0.112*	(0.067)	0.603***	(0.051)
Interviewed alone	0.233***	(0.063)	0.103*	(0.055)	0.599***	(0.080)	0.058	(0.064)
HH size	-0.064**	(0.030)	0.004	(0.026)	-0.035	(0.035)	0.009	(0.027)
Rent home	-0.050	(0.058)	0.016	(0.052)	-0.100	(0.074)	-0.131**	(0.063)
Married	-0.73***	(0.076)	-0.245***	(0.067)	-0.609***	(0.096)	-0.295***	(0.081)
Widowed	-0.034	(0.075)	-0.006	(0.072)	0.310***	(0.113)	0.081	(0.106)
White	0.118	(0.125)	0.045	(0.115)	-0.242	(0.153)	0.023	(0.129)
Retired	-0.018	(0.068)	0.016	(0.059)	-0.17	(0.213)	-0.195	(0.186)
Unemployed	0.623***	(0.235)	0.082	(0.188)	-0.484*	(0.254)	-0.462**	(0.221)
Working	-0.039	(0.075)	0.036	(0.061)	-0.273	(0.200)	-0.299*	(0.176)
SRH	✓		✓		✓		✓	
Age splines	✓		✓		✓		✓	
Smoking behavior	✓		✓		✓		✓	
Medical conditions	✓		✓		✓		✓	
Service use	✓		✓		✓		✓	
Family history	✓		✓		✓		✓	
Occupational class	✓		✓		✓		✓	
ρ	0.233***	(0.029)			0.149***	(0.035)		
N	4,648				3,556			

* p<0.10, ** p<0.05, *** p<0.01

Notes: Coefficients after bivariate probit analysis displayed. Standard errors in parentheses. a is equal to 1 if an individual is observed at least once after ELSA wave 0, h is equal to 1 if all objectively measured health variables are observed for an individual.

Chapter 2

Do they know what's at risk? Health Risk Perception Among the Overweight and Obese

2.1 Introduction

Excess body weight is a risk factor for various diseases. In particular, it increases the risks of type 2 diabetes, cardiovascular disease, several cancers, arthritis, and psychological problems (Haslam and James, 2005; Dixon, 2010). Today, more than 70% of US adults are overweight or obese, meaning that they have a Body Mass Index (BMI) between 25 and 30, or of at least 30, respectively.[1] Furthermore, a large share of US medical expenditures is caused by excess weight (Cawley and Meyerhoefer, 2010).

While the causes of excess body weight are multifactorial, body weight depends on each individual's behavior. Individuals can choose how many calories they consume and whether they engage in activities with high or low calorie expenditure, for example they can choose whether they are physically active.

Attempts to help individuals lose weight have shown limited success. In a meta-analysis McTigue et al. (2003) find that various interventions only promoted modest weight loss. One possible explanation for this is that individuals who carry excess weight are not aware of the danger that this weight poses to their health. Gregory et al. (2008) accordingly find that individuals who do not rate their weight as a health risk are less likely to engage in weight loss activities than individuals who rate their weight as a health risk.

[1]The BMI is measured as $\frac{weight(kg)}{height^2(m^2)}$. Information on overweight and obesity in the US from the National Health and Nutrition Examination Survey (NHANES), 2007/08.

In this chapter we analyze whether middle-aged individuals in the US with excess body weight underestimate their personal risks of diabetes, stroke, heart attack, lung disease, hypertension and arthritis or rheumatism. Health risk perceptions among overweight and obese individuals have been studied by Kan and Tsai (2004) and Gregory et al. (2008). Our analysis contributes to this literature by measuring individual risk perceptions in a different way. In particular, while Kan and Tsai (2004) and Gregory et al. (2008) use information on individual risk perception stemming from qualitative questions, we ask individuals about numerical probabilities of developing different diseases in the future. An advantage of our numerical subjective measures is that they can be compared to objective probabilities. This allows us to investigate how accurate individual risk perceptions are.

Our data on subjective disease probabilities comes from the American Life Panel (ALP), a panel study administered by RAND. In an ongoing survey in the ALP that went into the field in the summer of 2010 we ask the survey participants to assess their chances of developing different diseases in the next 5 years on a 0-100 scale. In the design of these health expectations questions we follow Manski (2004) who also provides a general assessment of the validity of similar expectation questions in different domains, such as stock market returns and returns to schooling.

In order to gauge how well individuals are informed about their health risks we estimate objective probabilities of developing the different diseases in the next 5 years. In the estimation of these objective disease probabilities we follow Khwaja et al. (2009) who conduct a similar analysis on risk perception among smokers. More specifically, we estimate relationships between individual characteristics and future disease onsets in the Health and Retirement Study (HRS). Under the assumption that these relationships are the same in the ALP today as they were in the HRS a few years ago, we use the estimates from the HRS to predict the probabilities of disease onsets for individuals in the ALP.

While not the main focus of our analysis, we use the subjective and objective disease probabilities in the ALP data to replicate the study by Khwaja et al. (2009) on the accuracy of risk perception among current and former smokers. The results are presented in the appendix to this chapter. Essentially, we are able to reproduce their findings that current and former smokers do not underestimate the different health risks.

Obese individuals in the ALP, on the contrary, underestimate certain health risks. In particular, we find evidence that obese individuals underestimate the risks of developing arthritis, diabetes and hypertension within the next 5 years. The results for

arthritis are robust to changes in the calculation of the objective risks. The risks of developing other diseases, such as a heart attack or a stroke, however, are significantly overestimated by the overweight and obese. As the objective risks of developing these diseases are relatively small on average, even among obese individuals, this overestimation corresponds to results from other studies that find that individuals tend to overestimate low probability events (see e.g. Lichtenstein et al., 1978).

The significant underestimation of the risks of arthritis, diabetes and hypertension among the obese indicates that there is a need for health education in this group. While health education programs targeted at the overweight and obese will probably not suffice to fully control the obesity epidemic, our results indicate that there is room for health education programs to help these individuals make better informed lifestyle choices.

The chapter is structured as follows. Section 2 summarizes the related literature. Section 3 describes the two data sets that we use in this analysis. In Section 4 the empirical methods are explained. Section 5 presents the results. Section 6 explores the robustness of the results, and Section 7 concludes.

2.2 Related Literature

Our analysis is closely related to two strands of literature. The first strand focuses directly on risk perceptions among the overweight and obese. The second strand analyzes risk perception among smokers.

The different studies that analyze risk perception among the overweight and obese focus primarily on the relationship between risk knowledge and the tendency to be overweight or obese rather than on the accuracy of the risk knowledge. Kan and Tsai (2004) find that there is a relationship between individuals' perception of health risks resulting from excess weight and the tendency to carry excess weight in Taiwanese men. In particular, men with very high BMI seem to be less aware of obesity-related health risks. For women, the authors find no relationship between health risk perception and BMI. Similarly, Gregory et al. (2008) find that overweight and obese adults in the US who do not perceive their weight as a risk factor are less likely to engage in weight lowering activities. Furthermore, the authors show that significant fractions of overweight and obese adults do not agree to the statement that their weight is a health risk. This suggests a lack of health knowledge on weight as a risk factor. However, the question that measures health knowledge is targeted at overall health risks and is thus very general. It does not allow to distinguish between risks of different diseases.

Moreover, because of the question's qualitative nature it is not clear how accurate individual risk perceptions are.

Analyses of risk perception among smokers have focused on the accuracy of the risk perception. Viscusi (1990) pioneered this type of analysis in economics by comparing individuals' assessments of the probability that smokers get lung cancer because they smoke with objective measures of this probability. His results indicate that smokers and non-smokers significantly overestimate the risk of lung cancer due to smoking. Viscusi and Hakes (2008) extend the earlier analysis to the accuracy of individuals' assessments of the probability that smokers die from lung cancer, heart disease, or any other illness because they smoke, and to the accuracy of loss in life expectancy due to smoking. In line with Viscusi's earlier findings their results indicate that smokers and non-smokers overestimate the risk of dying from smoking and the years of life lost due to smoking when compared to the risks of dying from smoking and life years lost as reported in the medical literature. As smokers overestimate these risks less than non-smokers, the results are consistent with the idea that higher risk perceptions protect individuals from smoking.

While Viscusi (1990) and Viscusi and Hakes (2008) study perceptions of general risks of smoking, i.e. individuals are asked about the risks of a hypothetical smoker, Schoenbaum (1997) and Khwaja et al. (2009) analyze the accuracy of risk estimates concerning the individuals' personal risks. Schoenbaum (1997) compares individual expectations of reaching age 75 measured in the HRS with actuarial predictions based on life tables for the groups of never smokers, former smokers, current light smokers, and current heavy smokers. Heavy smokers are found to significantly overestimate the probability of reaching age 75, while the other groups have about accurate risk perceptions. In Khwaja et al. (2009), individual expectations of reaching age 75 and expectations of getting certain diseases until the age of 75 are compared with individual objective risk estimates for each event. The objective risk estimates are based on the relationship between individual characteristics and disease probabilities as estimated from the HRS. Their results indicate that smokers do not underestimate disease and mortality risks.

2.3 Data

The analysis in this chapter is based on two data sets. The main analysis is conducted with data from the American Life Panel (ALP) in which we elicited subjective risk expectations. Data from the Health and Retirement Study (HRS) is used in the

calculation of objective health risks for individuals in the ALP.

The HRS was started in 1992 as a representative study of the US non-institutionalized population born between 1931 and 1941. Blacks, Hispanics and residents of Florida were oversampled. Sampling weights are provided to correct for this oversampling. Data was collected every two years since 1992. The last currently available data is from 2008. In later waves, additional cohorts were added to the HRS. In particular, in 1998 the "children of depression (CODA)" cohort (born 1924-1930) and the "war baby" cohort (born 1942-1947) were added to the data. We use the 1992 HRS wave and follow-up until 2008 to estimate prediction models for onset of diseases for a population aged 50-62. As a robustness check we also use a later HRS wave as a baseline year in the risk prediction models. In particular, we use the 1998 wave. The integration of additional cohorts in this wave allows to focus the analysis on a representative sample for the same age group (50-62 year-olds) as when the 1992 wave is used as baseline.

The second data source is the American Life Panel (ALP), an internet survey of about 3,200 American adults administered by RAND (see *http://rand.org/labor/roybalfd/american_life.html* for a full description). We implemented a survey in the ALP that went into the field in July 2010 and is still ongoing. The focus of our survey lies primarily on eliciting individuals' subjective expectations of developing certain diseases in the future. We elicit the expectations as numerical probabilities with the question "What do you think is the percent chance that you will develop, or re-develop if you have already been diagnosed with it, the following conditions in the next 5 years and ever in you lifetime?".[2] The question is preceded by a text that explains the probability scale from 0 to 100, where 0 means that there is absolutely no chance, or 0 percent, and 100 means the event is absolutely sure to happen, or 100 percent. Similar types of questions on survival expectations have proven useful in eliciting subjective probabilities that have predictive power for actual outcomes (see Hurd, 2009, for a summary).

In addition to the subjective expectations we elicited information on risk factors for the specific diseases, including questions on co-morbidities, family history, and lifestyle. 915 of the surveyed individuals in our sample are between 50 and 62 years old at the time of the survey. While the ALP is not a representative sample for the US population, RAND provides weights to make it comparable to the Current Population

[2]Individuals who report that they have been diagnosed with chronic diseases, such as diabetes or chronic lung disease, are not asked about their chances of developing the specific conditions. For conditions that may relapse, like a stroke or a heart attack, all individuals are asked about the chances independent of their prior history of the disease. In this chapter, we focus on the 5 year risks. In a follow-up analysis, we plan to also investigate the accuracy of the lifetime risk perceptions.

Survey. These weights are used in all following analyses.

Table 2.1: Descriptive Statistics

	HRS 1992		ALP	
Variable	Mean	SD	Mean	SD
Self-rated health				
Excellent	0.24	0.43	0.08	0.32
Very Good	0.30	0.47	0.41	0.61
Good	0.27	0.44	0.33	0.58
Fair	0.13	0.34	0.15	0.44
Poor	0.07	0.26	0.03	0.22
BMI				
Normal weight (BMI<25)	0.35	0.48	0.27	0.55
Overweight (25≤BMI<30)	0.41	0.49	0.36	0.59
Obese 1 (30≤BMI<35)	0.16	0.37	0.22	0.51
Obese 2 (35≤BMI<40)	0.04	0.20	0.07	0.31
Obese 3 (BMI≥40)	0.02	0.14	0.08	0.33
Smoking Status				
Current Smoker	0.27	0.44	0.21	0.50
Former Smoker	0.37	0.48	0.32	0.57
Never Smoker	0.36	0.48	0.47	0.62
Demographics				
Age	55.56	3.25	55.55	4.52
White	0.86	0.345	0.84	0.45
Male	0.48	0.50	0.49	0.62
Married	0.74	0.44	0.68	0.57
Education				
Less than High School	0.23	0.42	0.06	0.28
High School or equiv	0.39	0.49	0.33	0.58
Some College	0.20	0.40	0.30	0.56
BA or equiv	0.11	0.31	0.20	0.49
More than BA	0.08	0.27	0.12	0.40
N	9,793		906	

Notes: Sampling weights used in estimation of means and standard deviations (SD).

In table 2.1 weighted means and standard deviations of relevant health variables, risk factors and demographics are displayed for individuals aged 50 to 62 in the ALP data. For means of comparison, descriptives are also displayed for the HRS 1992 sample.

While the HRS and ALP samples seem to be similar in age and sex, there are differences in BMI categories, smoking status, and education. Less of the individuals

in the ALP sample are current or former smokers, but more of them are overweight or obese. These differences align with changes in the prevalence of smoking and obesity over the last decades. While the fraction of smokers decreased, the fraction of individuals with excess weight increased.[3] The individuals in the ALP 2010 are furthermore on average better educated than the individuals in the HRS in 1992. This could reflect the steady advance in educational accomplishments in the US.[4] For the results of our analysis the differences in individual characteristics between the samples do not matter as long as the relationship between the individual characteristics and the probabilities of developing different diseases in the future are constant over time. We shed light on the question how strong this assumption is in the aforementioned robustness check in which we use a later HRS wave as baseline.

The first two columns of table 2.2 display averages and standard errors of the subjective disease probabilities in the ALP for individuals who are between 50 and 62 years old. For each disease, only individuals are included who do not report ever having been diagnosed with the respective disease. In addition to the overall averages, means of the subjective probabilities are shown for the different BMI categories. It can be observed that the overall means vary between conditions. The means range from an 8 percent risk of developing lung disease in the next 5 years to a 21 percent risk of developing arthritis or rheumatism. Furthermore, the average risks vary by categories of BMI. For example, individuals with a BMI between 30 and 35 on average rate their risk of diabetes in the next 5 years more than 2 times as high as individuals with a BMI lower than 25. For none of the conditions, however, there is a clear graded relationship between average subjective risk and BMI.

2.4 Estimation Methods

In order to evaluate the accuracy of individuals' subjective beliefs on different risks we construct objective risks of getting different diseases in the next 5 years for every individual in the ALP. In a second step, we analyze whether the difference between the subjective and objective risk is related to the individuals' BMI.

[3]See National Center for Health Statistics (2010), page 24ff. for trends in smoking and obesity in the US.

[4]The trend in educational attainment can be seen in data from the U.S. Census available at *www.census.gov/hhes/socdemo/education/data/cps/historical/index.html*. The data shows that in 1992 25.7% of the 35 to 55 year old individuals and 14.2% of the individuals aged 55 or above in the US had finished their education with 4 years of college or more compared to 27.3% of the younger and 26.4% of the older group in 2009.

Table 2.2: Subjective and Objective Risk

		Subjective Risk		Objective Risk	
		Mean	SE	Mean	SE
Diabetes	Total	10.58	0.72	5.34	0.24
	BMI<25	7.14	1.04	1.22	0.05
	25≤BMI<30	10.45	1.09	3.89	0.08
	30≤BMI<35	15.15	2.04	8.79	0.25
	35≤BMI<30	10.48	1.64	12.76	0.69
	BMI≥40	14.44	3.81	17.62	0.58
Stroke	Total	12.37	0.81	1.91	0.06
	BMI<25	10.05	1.37	1.41	0.08
	25≤BMI<30	11.28	1.06	1.52	0.07
	30≤BMI<35	14.88	2.12	2.12	0.12
	35≤BMI<40	18.44	3.50	2.47	0.23
	BMI≥40	13.37	3.75	4.54	0.40
Lung Disease	Total	8.14	0.62	2.55	0.11
	BMI<25	7.71	1.21	2.39	0.23
	25≤BMI<30	8.70	1.18	2.19	0.14
	30≤BMI<35	9.00	1.19	2.70	0.24
	35≤BMI<40	6.22	1.32	3.37	0.41
	BMI≥40	5.95	1.18	3.84	0.39
Heart Attack	Total	13.07	0.79	3.13	0.11
	BMI<25	9.90	1.38	1.99	0.17
	25≤BMI<30	13.81	1.17	3.07	0.15
	30≤BMI<35	14.35	2.06	4.13	0.30
	35≤BMI<40	17.07	2.79	4.17	0.50
	BMI≥40	14.13	3.19	3.99	0.45
Hypertension	Total	11.53	0.94	12.80	0.23
	BMI<25	9.00	1.51	9.29	0.11
	25≤BMI<30	11.54	1.25	12.47	0.13
	30≤BMI<35	11.51	2.14	15.45	0.22
	35≤BMI<40	23.06	5.30	18.33	0.35
	BMI≥40	15.48	5.43	21.75	0.57
Arthritis	Total	21.14	1.37	21.06	0.40
	BMI<25	23.16	2.57	15.63	0.38
	25≤BMI<30	20.32	2.13	19.44	0.33
	30≤BMI<35	19.84	2.64	23.85	0.53
	35≤BMI<40	19.58	4.85	30.04	1.01
	BMI≥40	22.77	8.17	37.32	1.15

Notes: Individuals in the ALP aged 50-62 who have not been diagnosed with the respective disease before. Weights used in calculating means and standard errors (SE). Subjective expectation on developing of arthritis/rheumatism is combination of subjective expectation on arthritis/rheumatism onset and subjective expectation of rheumatoid arthritis onset (see appendix to this chapter for details).

2.4.1 Calculation of Objective Risk

For the calculation of the objective risks we use data from the HRS to estimate the relationship between individual characteristics at a baseline year and the observed onset of diseases in the future. Similar to Khwaja et al. (2009) we use d, $d \in \{1, ..., D\}$ duration models to model the relationship between the characteristics and the time to onset of each disease, d. We assume that the survivor functions follow Weibull distributions and allow for Gamma distributed unobserved heterogeneity. The survivor function for disease d is

$$S(t_i^d; \mathbf{X_i}, \boldsymbol{\theta}^d, \mu^d | \eta_i^d) = \left(exp(-\lambda_i^d (t_i^d)^{\mu^d}) \right)^{\eta_i^d} \tag{2.1}$$

where λ_i^d is parameterized as $exp(-\mu^d \mathbf{X_i}' \boldsymbol{\theta}^d)$ so that the survival distribution is a Weibull in accelerated failure time (AFT) metric. η_i^d stands for the unobserved heterogeneity that is Gamma($\frac{1}{\sigma^d}$,σ^d) distributed. t_i^d represents the time until individual i is diagnosed with disease d. If individual i does not report a diagnosis of the disease within the time that she is observed in the HRS, i is treated as a right-censored observation. In this case, t_i^d contains the time to censoring, i.e. the time in which i is observed.[5] μ^d is the shape parameter of the Weibull distribution, $\mathbf{X_i}$ represents the vector of individual characteristics and includes information on smoking status, BMI categories, self-rated health, age, sex, educational degree, marital status, and race. $\boldsymbol{\theta}^d$ is the corresponding vector of coefficients.

We estimate μ^d, $\boldsymbol{\theta^d}$, and σ^d by maximum likelihood. Based on these parameter estimates, we calculate each individual i's probability of not getting disease d in the next t years as

$$\widehat{O_i^d}(t; \mathbf{X_i}, \hat{\boldsymbol{\theta}}^d, \hat{\mu}^d) = E_\eta \left(\widehat{O_i^d}(t; \mathbf{X_i}, \hat{\boldsymbol{\theta}}^d, \hat{\mu}^d | \hat{\eta}^d) \right) = \left[1 + \hat{\sigma}^d exp(-\hat{\mu}^d \mathbf{X_i}' \hat{\boldsymbol{\theta}}^d) t^{\hat{\mu}^d} \right]^{-\frac{1}{\hat{\sigma}^d}} \tag{2.2}$$

Naturally, the probability of getting the disease within the next t years is then

$$\widetilde{O_i^d} = 1 - \widehat{O_i^d}(t; \mathbf{X_i}, \hat{\boldsymbol{\theta}}^d, \hat{\mu}^d). \tag{2.3}$$

As the time horizon of the subjective risk is 5 years, we set $t = 5$.

We have to make two main assumptions to use the estimated relationship for the

[5] The time to disease onset, t_i^d, is measured in years. A report of a new diagnosis in wave A is coded as the time in years between wave 1 and wave A minus 1. As there are two years between consecutive waves, t_i^d is coded as 5, for example, in the case of a new report of a disease d by individual i in wave 4. For right-censored observations t_i^d is measured in years between 1992 and the year of the observation i's last interview.

calculation of the objective risks in the ALP. As above-mentioned, we assume that the relationships between the characteristics and disease onsets stay constant over time and thus are the same in the ALP today as they were in the HRS. This assumption is evaluated by using different HRS waves as baseline years in robustness analyses.

Furthermore, we assume that the development of the different diseases is accurately captured by the reports of the diagnoses of the diseases in the HRS. This assumption might fail for at least two different reasons. On the one hand, individuals might be unaware of having a disease and therefore do not report it in the HRS. In particular, two of the diseases focused on in this study, diabetes and hypertension, are conditions that tend to remain undiagnosed, because they are often asymptomatic (Chatterji et al., 2010). On the other hand, individuals might misreport diseases for other reasons than unawareness, for example to justify non-participation in the labor market (Baker et al., 2004). Up to now, we do not further analyze the consequences that potential misreporting has for our results.

2.4.2 Comparison Subjective and Objective Risk

In the next step, we compare the subjective and the objective disease risks. Following Khwaja et al. (2009), we regress the difference between the subjective probability and the objective probability on the different BMI categories with normal- or underweight as reference

$$S_i^d - \widetilde{O_i^d} = \gamma_0^d + \gamma_1^d OV_i + \gamma_2^d OB1_i + \gamma_3^d OB2_i + \gamma_4^d OB3_i + \epsilon_i^d \tag{2.4}$$

where S_i^d indicates the subjective risk of developing disease d within the next 5 years, OV is a dummy for having a BMI between 25 and 30, and $OB1$, $OB2$, and $OB3$ stand for the 3 categories of obesity, $30 \leq$ BMI < 35, $35 \leq$ BMI < 40, and BMI≥ 40, respectively. If γ_0^d is significantly positive, individuals who are normal- or underweight significantly overestimate their risk of developing disease d. Significantly positive estimates for γ_2^d, γ_3^d or γ_4^d indicate that in the respective group the difference between subjective and objective risk is on average larger than among the normal- or underweight individuals. Individuals in group k on average underestimate their risk of disease d if $\gamma_0^d + \gamma_k^d$ is significantly smaller than 0.

In addition to these simple comparisons of means between the groups, we include information on income and education in equation (2.4) to analyze whether differences in these socio-economic characteristics can explain differences in the accuracy of risk perception between different BMI groups.

2.5 Results

Table 2.3 reports coefficients and standard errors after estimation of the model specified in equation (2.1) for the different diseases. As the models are specified in accelerated failure time metric, a positive coefficient indicates that the time to disease onset increases with the associated variable, i.e. the risk decreases with the variable. Conversely, a negative coefficient indicates that the risk increases with an increase in the associated variable. The risks of all diseases except diabetes and hypertension significantly increase with age. The disease risks are higher for worse assessments of self-rated health, and they increase with smoking and BMI categories. Lower educated individuals are at higher risk for diabetes, lung disease and hypertension, and men are at higher risk for diabetes, stroke and a heart attack, while women have a higher risk of arthritis or rheumatism.

Based on the estimates in table 2.3, objective disease risks for the individuals in the ALP are calculated according to equations (2.2) and (2.3). The last two columns of table 2.2 report the average objective probabilities and standard errors for the sample of 50-62 year-olds in the ALP. Averages of objective probabilities are also displayed for the different categories of BMI. The average objective risks show a clear graded relationship with BMI. All disease risks are on average higher for higher BMI categories.

Table 2.4 reports results of the estimation of equation (2.4) for the different diseases. The first column of results for each disease reports estimates when only indicators for the different BMI categories are included as explanatory variables. In the reference category, i.e. among individuals who have a BMI smaller than 25, the subjective probabilities of all disease onsets but hypertension are on average significantly larger than the respective objective probabilities. Normal- and underweight individuals, for example, overestimate the risk of being diagnosed with diabetes within the next 5 years by 5.9 percentage points on average. For the other diseases the overestimation ranges from 5.3 percentage points for lung disease to 8.6 percentage points for stroke.

The results for the overweight and obese individuals are mixed for the different diseases. Similar to the normal- and underweight, they seem to correctly asses their risk of hypertension and overestimate the risks of a stroke, a heart attack, and the onset of chronic lung disease within the next 5 years. The risk of diabetes is also overestimated by the overweight and mildly obese ($30 \leq BMI < 35$). The two highest BMI categories, however, seem to assess their risk accurately on average. In these two categories, the sum of the group's coefficient and the constant is not significantly different from 0.

Table 2.3: Duration Model for Disease Onsets in HRS

	Diabetes		Stroke		Lung Disease		Heart Attack		Hypertension		Arthritis	
Age	-0.008	(0.006)	-0.051***	(0.009)	-0.022***	(0.007)	-0.033***	(0.011)	-0.004	(0.004)	-0.015***	(0.005)
Male	-0.185***	(0.043)	-0.131**	(0.061)	0.048	(0.051)	-0.589***	(0.078)	0.039	(0.026)	0.307***	(0.035)
SRH – very good	-0.103*	(0.060)	-0.165*	(0.090)	-0.320***	(0.087)	-0.421***	(0.110)	-0.025	(0.033)	-0.257***	(0.043)
SRH – good	-0.215***	(0.062)	-0.459***	(0.089)	-0.554***	(0.087)	-0.713***	(0.111)	-0.067*	(0.035)	-0.407***	(0.049)
SRH – fair	-0.359***	(0.074)	-0.700***	(0.111)	-0.829***	(0.097)	-1.155***	(0.132)	-0.109**	(0.047)	-0.418***	(0.067)
SRH – poor	-0.491***	(0.097)	-0.943***	(0.130)	-0.973***	(0.114)	-1.290***	(0.171)	-0.113*	(0.064)	-0.433***	(0.094)
White	0.180***	(0.051)	0.075	(0.075)	-0.356***	(0.068)	-0.282***	(0.093)	0.183***	(0.032)	-0.026	(0.043)
Married	0.047	(0.047)	0.087	(0.066)	0.054	(0.057)	-0.107	(0.086)	-0.038	(0.030)	-0.104**	(0.041)
Less than HS	-0.181***	(0.053)	0.038	(0.076)	-0.110*	(0.061)	-0.097	(0.091)	-0.049	(0.032)	-0.001	(0.046)
Some College	-0.025	(0.057)	-0.008	(0.080)	-0.048	(0.070)	-0.036	(0.094)	0.057	(0.035)	0.030	(0.045)
BA or eq.	-0.041	(0.076)	0.024	(0.105)	0.199*	(0.111)	-0.095	(0.133)	0.047	(0.045)	0.079	(0.057)
More than BA	0.095	(0.087)	0.183	(0.134)	-0.025	(0.117)	0.177	(0.158)	0.099*	(0.051)	-0.003	(0.065)
Current Smoker	-0.145***	(0.052)	-0.418***	(0.076)	-1.304***	(0.079)	-0.792***	(0.097)	-0.024	(0.032)	-0.090**	(0.044)
Former Smoker	-0.001	(0.048)	-0.149	(0.074)	-0.562***	(0.077)	-0.219**	(0.086)	0.009	(0.030)	-0.113***	(0.039)
25≤BMI<30	-0.736***	(0.060)	-0.019	(0.068)	-0.005	(0.059)	-0.318***	(0.087)	-0.210***	(0.029)	-0.185***	(0.039)
30≤BMI<35	-1.241***	(0.072)	-0.170*	(0.089)	-0.108	(0.073)	-0.519***	(0.107)	-0.351***	(0.038)	-0.371***	(0.050)
35≤BMI<40	-1.488***	(0.094)	-0.171	(0.138)	-0.190*	(0.115)	-0.510***	(0.183)	-0.471***	(0.060)	-0.556***	(0.094)
BMI≥40	-1.705***	(0.153)	-0.493**	(0.194)	-0.222	(0.149)	-0.451**	(0.229)	-0.546***	(0.111)	-0.698***	(0.156)
Constant	4.922***	(0.363)	7.294***	(0.538)	7.003***	(0.449)	7.765***	(0.641)	3.258***	(0.206)	4.080***	(0.272)
μ	1.617	(0.059)	1.764	(0.105)	1.318	(0.041)	1.418	(0.069)	1.474	(0.025)	1.260	(0.032)
σ	1.121	(0.357)	4.907	(1.768)	0.000002	(0.000003)	4.438	(1.399)	0.0000001	(0.00000002)	0.450	(0.165)
N	8,850		9,552		9,876		9,223		6,825		6,872	

* p<0.10, ** p<0.05, *** p<0.01

Notes: Robust standard errors in parentheses. Weibull AFT parameterization with Gamma distributed heterogeneity. μ shape parameter of Weibull, σ parameter of Gamma distribution. Small values of $\widehat{\sigma}$ indicate negligible heterogeneity. In this case, models without frailty would deliver almost identical results. Estimates based on HRS 1992 cohort aged 50-62 and follow up until 2008. Each estimation only includes individuals who do not report having been diagnosed with specific disease before 1992. Results are weighted using sampling weights provided with the data.

Table 2.4: Differences Subjective and Objective Risk

	Diabetes		Stroke		Lung Disease		Heart Attack		Hypertension		Arthritis	
Overweight (25≤BMI<30)	0.006	0.006	0.011	0.01	0.012	0.011	0.028	0.029*	-0.006	-0.009	-0.067**	-0.066*
	(0.015)	(0.014)	(0.017)	(0.016)	(0.017)	(0.016)	(0.018)	(0.017)	(0.019)	(0.018)	(0.033)	(0.034)
Obese 1 (30≤BMI<35)	0.004	0.006	0.041	0.034	0.01	0.007	0.023	0.02	-0.036	-0.032	-0.115***	-0.109***
	(0.023)	(0.021)	(0.025)	(0.025)	(0.017)	(0.017)	(0.025)	(0.023)	(0.027)	(0.022)	(0.037)	(0.038)
Obese 2 (35≤BMI<40)	-0.082***	-0.083***	0.073**	0.07**	-0.025	-0.029	0.05	0.053*	0.05	0.061	-0.18***	-0.191***
	(0.022)	(0.021)	(0.037)	(0.035)	(0.017)	(0.018)	(0.032)	(0.031)	(0.055)	(0.048)	(0.054)	(0.06)
Obese 3 (BMI≥40)	-0.091**	-0.09**	0.002	-0.007	-0.032*	-0.036*	0.022	0.019	-0.06	-0.051	-0.221***	-0.212***
	(0.041)	(0.035)	(0.041)	(0.039)	(0.017)	(0.019)	(0.036)	(0.03)	(0.056)	(0.051)	(0.081)	(0.072)
Constant	0.059***	0.072***	0.086***	0.059***	0.053***	0.024*	0.079***	0.049**	-0.003	0.026	0.075***	0.019
	(0.011)	(0.021)	(0.014)	(0.014)	(0.013)	(0.013)	(0.014)	(0.013)	(0.015)	(0.025)	(0.026)	(0.03)
Education	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Income	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	779	779	861	860	828	827	861	860	536	535	585	584
R^2	0.0333	0.0897	0.0148	0.0537	0.0097	0.056	0.0072	0.0562	0.0187	0.17	0.0511	0.088

* p<0.10, ** p<0.05, *** p<0.01

Notes: Standard errors in parentheses. Each estimation is based on individuals in the ALP aged 50-62 who do not report having been diagnosed with the respective condition. Results weighted using ALP sampling weights. Subjective expectation on the onset of arthritis is combination of subjective expectation on onset of arthritis/rheumatism and subjective expectation on onset of rheumatoid arthritis (see appendix to this chapter for details). Number of observations drops by 1 with inclusion of income for all conditions but diabetes, as there is one individual that has diabetes but no other disease and missing income information.

As the objective risks of most diseases for the normal- and underweight and of chronic lung disease, heart attack and stroke among the overweight and obese are relatively small on average, the general pattern of overestimation is in line with other results in the literature according to which individuals tend to overestimate small probability events (Lichtenstein et al., 1978). As diabetes risk increases significantly for the higher BMI groups, this might also explain why obese individuals have on average more accurate assessments of their risks of diabetes than the normal- and underweight.

Interestingly the results in table 2.4 suggest that obese individuals underestimate one health risk, namely the risk of developing arthritis or rheumatism in the next 5 years. On average, individuals with a BMI between 35 and 40 underestimate the risk by 10 percentage points and individuals in the highest BMI group underestimate the risk by about 15 percentage points.

The coefficients in the second column for each disease in table 2.4 result from estimation of equation (2.4) with information on an individual's educational attainment and on household income included in addition to indicators of BMI categories. The results are essentially unchanged. Neither differences in educational attainment nor differences in income can thus explain the underestimation of arthritis risk among the obese.

2.6 Robustness Analyses

In this section, we present several robustness analyses. First, we use estimates based on a later HRS wave to calculate the objective risks in the ALP. Second, we employ probit models instead of duration models to estimate the relationship between individual characteristics and disease onset in the HRS. Third, we report estimates using the relative difference between subjective and objective risk instead of the absolute difference as dependent variable.

The upper panel of table 2.5 displays coefficients and standard errors after estimation of equation (2.4) where the objective risk measure is based on HRS 1998 and follow-up instead of on the first HRS wave.[6] In order to ensure comparability to the main analysis, only individuals aged 50-62 from the HRS 1998 are used to estimate model (2.1). Using a later cohort in the HRS as baseline helps us to investigate whether

[6]The latest available wave that allows 5 years follow-up is wave 6, elicited in 2002. When information from this wave and follow-up is used, results are qualitatively and quantitatively unchanged from the ones displayed in the upper panel of table 2.5.

Table 2.5: Robustness Analyses I

	Diabetes	Stroke	Lung Disease	Heart Attack	Hypertension	Arthritis
Objective Risk based on duration model with HRS 1998 and follow-up						
Overweight (25≤BMI<30)	-0.005	0.009	0.014	0.034*	-0.049**	-0.042
	(0.015)	(0.017)	(0.017)	(0.018)	(0.02)	(0.033)
Obese 1 (30≤BMI<35)	-0.022	0.046*	0.01	0.033	-0.075***	-0.139***
	(0.023)	(0.025)	(0.018)	(0.024)	(0.028)	(0.037)
Obese 2 (35≤BMI<40)	-0.134***	0.059	-0.025	0.056*	0.008	-0.193***
	(0.023)	(0.037)	(0.018)	(0.033)	(0.053)	(0.054)
Obese 3 (BMI≥40)	-0.128***	0.007	-0.055***	0.02	-0.062	-0.236***
	(0.041)	(0.041)	(0.018)	(0.036)	(0.053)	(0.083)
Constant	0.055***	0.083***	0.048***	0.076***	-0.041***	0.053**
	(0.01)	(0.014)	(0.013)	(0.014)	(0.015)	(0.026)
N	779	861	828	861	536	585
R^2	0.0643	0.0136	0.018	0.0101	0.0326	0.0668
Objective Risk based on probit model with HRS 1992 and follow-up						
Overweight (25≤BMI<30)	-0.002	0.014	0.018	0.03	-0.035*	-0.067**
	(0.015)	(0.017)	(0.017)	(0.018)	(0.019)	(0.033)
Obese 1 (30≤BMI<35)	-0.028	0.04	0.007	0.015	-0.078***	-0.114***
	(0.023)	(0.025)	(0.017)	(0.025)	(0.027)	(0.037)
Obese 2 (35≤BMI<40)	-0.128***	0.078**	-0.027	0.042	0.02	-0.221***
	(0.027)	(0.037)	(0.018)	(0.032)	(0.055)	(0.057)
Obese 3 (BMI≥40)	-0.177***	0.007	-0.014	0.029	-0.186***	-0.332***
	(0.043)	(0.041)	(0.017)	(0.036)	(0.06)	(0.084)
Constant	0.06***	0.083***	0.046***	0.073***	-0.011	0.037
	(0.011)	(0.013)	(0.013)	(0.014)	(0.015)	(0.026)
N	779	861	828	861	536	585
R^2	0.083	0.0151	0.0089	0.0067	0.0789	0.0899

* p<0.10, ** p<0.05, *** p<0.01

Notes: Standard errors in parentheses. Dependent variables are differences between individual subjective and objective risks. Each estimation is based on individuals in the ALP aged 50-62 without a diagnosis of the specific condition prior to the ALP interview. Results weighted to take into account sample selection. Subjective expectation on the onset of arthritis is combination of subjective expectation on arthritis/rheumatism onset and subjective expectation of rheumatoid arthritis onset (see appendix to this chapter for details).

and how changes in the relationship between individual characteristics and disease onset over time affect our results. Two important differences compared to the main results emerge. First, obese individuals do not only underestimate the risk of arthritis, but also the risk of diabetes. Second, the risk of being diagnosed with hypertension in the next 5 years is significantly underestimated among all BMI categories on average. The underestimation of the risk of hypertension is especially pronounced among the overweight and mildly obese.

As the measures of disease onset in the HRS rely on self-reports instead of for example on results from a medical examination, these results could reflect a general trend in the reduction of the number of undiagnosed cases of diabetes and hypertension over time.[7] Furthermore, an increased awareness of the health problems related to obesity might have made diagnoses of diabetes and hypertension particularly more likely among the higher BMI group. This could explain why the calculated objective risks of hypertension and diabetes are higher for the higher BMI categories when they are based on later HRS waves.

Table 2.6: Robustness Analyses II

	Diabetes	Stroke	Lung Disease	Heart Attack	Hypertension	Arthritis
Overweight (25≤BMI<30)	-3.265***	0.401	0.479	-2.112	-0.050	-0.459**
	(0.884)	(1.752)	(1.184)	(1.445)	(0.184)	(0.216)
Obese 1 (30≤BMI<35)	-4.260***	-0.621	0.493	-3.834***	-0.190	-0.684***
	(0.875)	(1.912)	(1.338)	(1.452)	(0.218)	(0.211)
Obese 2 (35≤BMI<40)	-5.199***	1.799	-2.216**	-2.000	0.314	-0.869***
	(0.851)	(3.327)	(1.058)	(2.029)	(0.332)	(0.239)
Obese 3 (BMI≥40)	-5.182***	-3.525	-2.733***	-2.569	-0.274	-0.929***
	(0.872)	(2.994)	(1.008)	(2.112)	(0.279)	(0.273)
Constant	5.058***	7.864***	3.564***	7.231***	-0.036	0.509***
	(0.840)	(1.505)	(0.922)	(1.318)	(0.159)	(0.180)
N	779	861	828	861	536	585
R^2	0.0779	0.006	0.01	0.0126	0.0084	0.043

* p<0.10, ** p<0.05, *** p<0.01

Notes: Standard errors in parentheses. Dependent variables are differences between individual subjective and objective risk relative to the objective risk of each disease. Each estimation is based on individuals in ALP aged 50-62 without respective condition at time of interview. Results weighted to take into account sample selection. Subjective expectation on the onset of arthritis is combination of subjective expectation on arthritis/rheumatism onset and subjective expectation of rheumatoid arthritis onset (see appendix to this chapter for details).

Results using objective risk estimates based on probit models instead of duration

[7]See Smith (2007) for the development of undiagnosed diabetes over time and Cutler et al. (2008b) for the development of undiagnosed hypertension.

models are presented in the lower panel of table 2.5. Instead of the time to disease onset as the dependent variable, we use an indicator variable that takes on the value one if an individual is diagnosed with the specific condition sometime before the 4th wave of the HRS and is zero otherwise. The results confirm conclusions drawn in the main analysis. Obese individuals underestimate the risk of arthritis or rheumatism. Moreover, there is evidence that the obese are overly optimistic about their risks of diabetes and hypertension, as they also underestimate the risks of the latter two diseases.

Table 2.6 reports results when the deviation of the subjective risk from the objective risk is measured relative to the objective risk. Instead of the difference in levels, we use the difference relative to the objective risk, $\frac{S_i^d - \widetilde{O_i^d}}{\widetilde{O_i^d}}$, as dependent variable in model (2.4). The results are again similar to the results of the main analysis. Individuals who are normal- or underweight overestimate the risks of most diseases significantly. As the percent rather than percentage point metric puts an emphasis on deviations from smaller objective values, it comes at no surprise that there are overestimations between 350 and almost 800% for the diseases that show relatively low risks among normal- and underweight individuals. The risk of hypertension is not overestimated and the risk of arthritis is overestimated by about 50% among the normal- and underweight. Importantly, despite putting relatively less emphasis on deviations from higher objective risks, we find evidence that individuals who are obese underestimate their risk of arthritis or rheumatism. Individuals with a BMI between 35 and 40 underestimate the risk by about 36% and individuals with a BMI over 40 underestimate their risk by 42%.

Overall, the robustness analyses confirm our main findings that obese individuals underestimate the 5-year risk of arthritis or rheumatism. Furthermore, this section provides evidence suggesting that also the risks of hypertension and diabetes are underestimated by individuals with excess body weight.

2.7 Conclusion

In this chapter, we add to the literature on risk perception among the overweight and obese by comparing subjective probabilities of developing different diseases within the following 5 years to objective probabilities of the same events. Our results indicate that individuals who are normal- or underweight overestimate most disease risks. Similarly, the overweight tend to overestimate their risks. Obese individuals, however, underestimate the risks of arthritis or rheumatism, diabetes and hypertension, while they overestimate the risks of heart attack and stroke.

These results are important for at least two reasons. First, diabetes and hypertension are silent conditions that often do not show symptoms. Therefore they risk to remain undiagnosed. When undiagnosed the conditions cannot be treated adequately and are thus likely more harmful. Individuals who underestimate the risks of these conditions might also be less likely to get checked for the conditions. The conditions might therefore remain undiagnosed longer, causing greater harm and higher medical expenses than if they were treated correctly.

Second, the results can provide guidance for designing health education programs targeted at the obese. While the medium term risks of cardiovascular disease, and in particular heart attack and stroke, is on average not underestimated by the obese, there is evidence that the risks of diabetes, hypertension and arthritis are not well understood. Increasing the awareness of these risks among the obese might allow them to make better informed lifestyle choices and help them to opt for a more healthy lifestyle.

Appendix

Results for smokers

In order to compare our results directly with the results of Khwaja et al. (2009), we estimate an equation similar to equation (2.4) where instead of the dummies for the BMI categories dummies for former and current smokers are included. Results are displayed in table 2.7. Despite some differences in the estimation method, the time horizon and in the data used, our results support the findings by Khwaja et al. (2009). Neither current nor former smokers robustly underestimate any of the analyzed disease risks.

Subjective Expectations on Arthritis/Rheumatism

Subjective expectations on developing arthritis or rheumatism in the future are elicited by two separate questions in the ALP. The first question asks individuals about their chances of developing arthritis or rheumatism except rheumatoid arthritis, $Pr(arthr)$, and the second question asks about the chances of developing rheumatoid arthritis, $Pr(RA)$. In the HRS, however, there is only information on arthritis and rheumatism including rheumatoid arthritis. We therefore combine the answers to the two subjective expectations question in the ALP.

As we are interested in whether overweight and obese individuals underestimate their risks of developing different diseases, we aggregate the two probabilities in a way that results in the largest possible subjective probability of developing arthritis. In particular, we set

$$Pr(arthritis) = min\{100; Pr(arthr) + Pr(RA)\} \tag{2.5}$$

Table 2.7: Results for Smokers

	Diabetes		Stroke		Lung Disease		Heart Attack		Hypertension		Arthritis	
					Objective Risk based on HRS 1992							
Current Smoker	-0.015	-0.012	-0.013	-0.015	0.022	0.026	-0.003	-0.009	-0.027	-0.00001	-0.014	-0.007
	(0.02)	(0.021)	(0.02)	(0.02)	(0.021)	(0.02)	(0.021)	(0.02)	(0.023)	(0.022)	(0.04)	(0.042)
Former Smoker	0.001	-0.003	0.011	0.003	0.012	0.006	0.004	-0.001	-0.003	-0.01	0.005	0.007
	(0.016)	(0.016)	(0.019)	(0.018)	(0.013)	(0.012)	(0.018)	(0.017)	(0.021)	(0.02)	(0.032)	(0.032)
Constant	0.056***	0.072***	0.104***	0.071***	0.048***	0.024***	0.099***	0.069***	-0.006	0.024	0.002	-0.038*
	(0.011)	(0.017)	(0.011)	(0.012)	(0.007)	(0.008)	(0.011)	(0.012)	(0.013)	(0.022)	(0.019)	(0.023)
Education	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Income	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	779	779	861	860	828	827	861	860	536	535	585	584
R^2	0.0016	0.0586	0.0023	0.0425	0.0042	0.0499	0.0002	0.049	0.0041	0.1537	0.0006	0.0424
					Objective Risk based on HRS 1998							
Current Smoker	-0.007	-0.002	-0.018	-0.02	0.003	0.011	0.005	0.0004	-0.03	0.0005	-0.034	-0.018
	(0.02)	(0.022)	(0.02)	(0.02)	(0.022)	(0.02)	(0.022)	(0.02)	(0.022)	(0.022)	(0.042)	(0.042)
Former Smoker	0.008	0.005	0.011	0.004	0.009	0.005	0.008	0.004	-0.003	-0.007	0.002	0.011
	(0.016)	(0.016)	(0.019)	(0.018)	(0.013)	(0.013)	(0.018)	(0.017)	(0.022)	(0.02)	(0.033)	(0.032)
Constant	0.032***	0.051***	0.101***	0.068***	0.047***	0.031***	0.097***	0.06***	-0.068***	-0.024	-0.012	-0.037
	(0.011)	(0.017)	(0.011)	(0.012)	(0.007)	(0.008)	(0.011)	(0.012)	(0.014)	(0.023)	(0.019)	(0.023)
Education	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Income	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	779	779	861	860	828	827	861	860	536	535	585	584
R^2	0.0012	0.0583	0.0034	0.0431	0.0009	0.0473	0.0005	0.0482	0.0048	0.1678	0.0026	0.0548

* p<0.10, ** p<0.05, *** p<0.01

Notes: Standard errors in parentheses. Each estimation is based on individuals in the ALP aged 50-62 who do not report having been diagnosed with the respective condition. Results weighted using ALP sampling weights. Subjective expectation on the onset of arthritis is combination of subjective expectation on onset of arthritis/rheumatism and subjective expectation on onset of rheumatoid arthritis (see appendix to this chapter for details). Number of observations drops by 1 with inclusion of income for all conditions but diabetes, as there is one individual that has diabetes but no other disease and missing income information.

Chapter 3

Heterogeneous Effects of Copayments: A Finite Mixture Bivariate Probit Analysis

3.1 Introduction

Analyses of reforms have traditionally focused on average effects. Average effects are helpful, for example, in evaluating the cost-effectiveness of a reform. A comparison of the average returns generated by a reform to the average costs of the reform suffices to draw conclusions on cost-effectiveness. The average effect, however, might hide details that are important for an overall evaluation of a reform. When the average effect is positive, for example, but it is a combination of positive effects for most individuals and large negative effects for a few, this is likely of interest to the policy maker. Depending on the policy maker's objective function, the harm to the few might be more important than the gains for the many. It is therefore important to analyze whether a reform affected different individuals differently.

In this paper, we develop and apply a finite mixture bivariate probit model that allows to simultaneously analyze two types of effect heterogeneities. On the one hand, the finite mixture component of the model allows to estimate different reform effects for different latent classes of individuals. More specifically, it is assumed that there are different classes or groups of individuals in the population, but that it is unobserved which individual belongs to which group. The observations within each group are assumed to be generated by the same distribution function. Between groups, however, any parameters of the distribution are allowed to vary. In the context of evaluating a reform, the model allows for as many distinct reform effects as there are groups of

individuals; within each group individuals are assumed to be affected in the same way.

On the other hand, the bivariate probit component of the model allows to jointly estimate the effect on two distinct binary outcomes. The combination of the finite mixture and the bivariate probit component permits to study effects on the two binary outcomes within each of the latent classes at the same time. To our best knowledge, these two heterogeneities have up to now only been investigated under the assumption of independence between the errors of the two outcome equations (Atella et al., 2004). In our model, this independence assumption is relaxed.

We apply the model to analyze heterogeneities in the effect of an increase of copayments for doctor visits and prescription drugs in the German statutory health insurance in 2004. The most radical element of the reform was an introduction of a per-quarter fee for doctor visits. Since 2004 patients have to pay €10 for the first visit to a physician in each quarter of the year. Further doctor visits to the same doctor within the quarter are free of charge for the patient. Visits in the same quarter to other doctors are also exempt from the fee if the patient gets a referral by the doctor whom he visited at first.

The literature that focuses on the effect of this specific reform delivers mixed results. Augurzky et al. (2006) and Schreyögg and Grabka (2010) find that the reform had essentially no effect on the health care use of the statutorily insured in the German Socio-Economic Panel (SOEP). Using the same data set Farbmacher (2010), on the contrary, presents evidence according to which the reform had an effect on the health care use of the statutorily insured. Farbmacher's results are in line with Rückert et al. (2008) who find that individuals surveyed in the Bertelsmann Healthcare Monitor report avoiding and delaying doctor visits due to the per-quarter fee. Our contributions to this literature are twofold. We use a new data set to study the effect of the reform. In addition, our study is the first that analyzes whether this reform had heterogeneous effects.

The data set that is employed in this analysis stems from insurance claims of the largest German sickness fund. It includes information on doctor visits two years before and two years after the reform. An advantage of using insurance claims data is that doctor visits are reliably observed. We observe every doctor visit irrespective of whether a patient paid the per-quarter fee. Furthermore, the data includes information on different types of doctor visits. In particular, it comprises information on visits to general practitioners (GPs) and on visits to specialists. This allows us to analyze whether the reform had differential effects on these two types of doctor visits. A disadvantage of the data is, however, that besides age and sex it does not contain any

socio-demographic information.

Furthermore, the data set only includes few individuals who are exempt from the copayments and could thus serve as a control group. We therefore apply a before-after comparison. Our results rely on the assumption that in the absence of the reform there would not have been any changes in health care use. In the two years before and two years after the introduction of the copayments there were no other reforms in the German health care system.

In our analysis, we focus on the question whether the increase of the copayments in 2004 had an effect on the probability to visit a GP and/or a specialist at least once within a year. We focus on the probability of visiting a doctor instead of on the number of doctor visits within that period, because the reform should mainly affect access to health care. The €10 fee only has to be paid for the first visit in a quarter. The price for additional doctor visits in the same quarter is therefore essentially zero.

Our results indicate that the average probability of at least one doctor visit, irrespective of the doctor's type, is 3 percentage points lower in the years after the reform than in the years before the reform. The results are in line with Rückert et al. (2008) and Farbmacher (2010) and indicate that the reform affected the access to health care.

Furthermore, we find evidence for both of the analyzed heterogeneities. When allowing for two latent classes, the reform has a much stronger effect on individuals in one of the latent classes. The more affected class only comprises 24% of individuals. The two groups can be characterized by their probability of health care use prior to the reform. The larger group has a probability of seeing at least one doctor per year that is almost 100%, we thus call them the *likely users*. The smaller group only goes to see a doctor with a probability of 66%. We call them the *less-likely users*. As the *likely users* react less to the reform than the *less-likely users* our results are in line with findings by Bago d'Uva (2006) and Winkelmann (2006) who present evidence that high users of health care are less sensitive to increases in coinsurance and copayments than low users.

In addition, we find evidence that the two groups react differently for the two types of doctor visits. Among the *likely users* the probability of seeing a GP is not affected by the reform. The probability of seeing a specialist, however, decreases after the reform in this group. For the *less-likely users*, on the contrary, the probabilities of both types of doctor visits decline significantly after the reform. For a large group of individuals the introduction of the per-quarter fee might thus have strengthened the GP's function as a gatekeeper in the health care system, even though the reform did not force individuals to focus on GP visits only. Our results thus provide evidence that

it might be possible to strengthen the GP's function as a gatekeeper without actually limiting individuals' choices directly.

The chapter is structured as follows: Section 2 introduces our finite mixture bivariate probit model, the data used in the analysis is described in Section 3. Section 4 presents the results, and Section 5 concludes.

3.2 Econometric Framework

Our data consists of a panel of individuals across time and across different physician types. The patient can seek care from GPs (y_{1it}) and/or specialists (y_{2it}). The panel is unbalanced over time, and each individual i is observed in T_i periods. Over time and physician types, individual i is thus observed $2 \cdot T_i$. Suppose that individual i belongs to a latent class j for the entire observational period. The probability of belonging to class j is π_j. Within a latent class, we use bivariate probits to jointly model the decision to visit a GP and/or a specialist. The joint probability of the dependent variables over the observed period is the product of T_i independent probabilities, given fixed class membership, i.e.,

$$Pr(y_{1i}, y_{2i}|x_i, \theta_j) = \prod_{t=1}^{T_i} \Phi_2[(2y_{1it} - 1)x_{it}\beta_j, (2y_{2it} - 1)x_{it}\gamma_j, (2y_{1it} - 1)(2y_{2it} - 1)\rho_j] \quad (3.1)$$

where x_i denotes the vector of covariates. θ_j contains the vector of parameters for GP visits (β_j), for specialist visits (γ_j), and in addition the parameter ρ_j. The latter parameter indicates the extent to which the errors in the underlying structural model covary.

The log-likelihood function is given by

$$ln(L) = \sum_{i=1}^{I} ln \left(\sum_{j=1}^{J} \pi_j Pr(y_{1i}, y_{2i}|x_i, \theta_j) \right) \quad (3.2)$$

where I is the number of individuals in the dataset and J is the number of latent classes. The likelihood function is maximized using the Newton-Raphson algorithm. First and second derivatives are calculated numerically in Stata's optimization package.

Our finite mixture bivariate probit model is similar to the model developed by Atella et al. (2004). We expand their model by a panel dimension and thus use additional information to determine class membership. While Atella et al. (2004) impose a correlation of zero between the errors of the two outcome equations, we estimate these

correlations for each latent class. As the dependent variables are visits to different types of physicians, the latter extension seems to be of particular importance.

For the interpretation of our results we calculate marginal effects on the marginal and joint outcome probabilities. Standard errors of the marginal effects are calculated using the delta method.[1]

3.3 Data

The analysis is based on insurance claims data from the largest German sickness fund in the years 2002 to 2005. The original data contains information on a 18.75% random subsample of all individuals in the German state of Hesse who are insured with this sickness fund.[2] We only include individuals in our analysis who are at least 19 years old, as younger individuals are exempt from the copayments.[3] Our sample is further restricted to all individual-year pairs for which we observe the use of medical services within the entire year. Individual-year pairs, for example, in years within which an individual switches from or to a different insurer or dies are excluded. This gives us the same period at risk for each observation. Furthermore, in order to get a manageable data set for our statistical analyses, we restrict the sample to a 3% random subsample. This gives us a sample with on average a little more than 7,000 individuals per year.

The data contains information on age and sex in addition to the information on health care use. Table 3.1 shows descriptive statistics for the different years. The average age is almost unchanged over time. This reflects that we use an unbalanced panel of individuals. The average number of doctor visits in our sample is almost 19 per year. On average individuals visit a GP a little less than once a month and a specialist around 7 times per year. This in international comparison relatively high use of physician services is in line with information from other data on doctor visits in Germany. The average number of doctor visits in another large German wide sickness fund, for example, is only slightly lower with 16.4 in 2004 and 16.9 in 2005 (Grobe et al., 2010).

While the average number of doctor visits per year does not change after the reform, at least one possible effect of the increased copayments becomes evident in table 3.1: Between 2003 and 2004, the fraction of individuals with at least one GP visit and the

[1] See appendix for details.

[2] See *http://www.pmvforschungsgruppe.de/content/02_forschung/02_b_sekundaerd_1.htm* for a short description of the data in German.

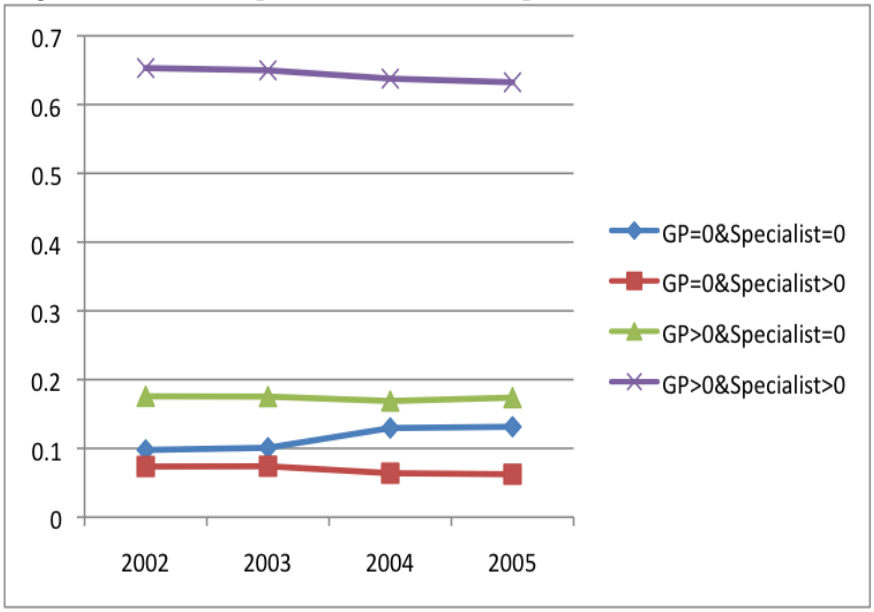
[3] This feature of the reform suggests a natural devision in treatment and control group. We conduct a difference-in-difference analysis for teenagers in a follow-up analysis.

Table 3.1: Descriptive Statistics

	2002		2003		2004		2005	
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age	52.35	18.55	52.47	18.66	52.58	18.64	52.49	18.72
19-39	0.30	0.46	0.29	0.45	0.28	0.45	0.28	0.45
40-59	0.31	0.46	0.32	0.46	0.32	0.47	0.33	0.47
60-79	0.32	0.47	0.32	0.47	0.32	0.47	0.31	0.46
$\geq$ 80	0.07	0.26	0.07	0.26	0.08	0.26	0.07	0.26
Female	0.52	0.50	0.52	0.50	0.52	0.50	0.52	0.50
# GP visits	11.56	13.87	11.42	14.22	11.71	14.61	11.90	15.56
GP>0	0.83	0.38	0.83	0.38	0.81	0.40	0.81	0.40
# GP visits truncated at 0	13.95	14.10	13.84	14.54	14.52	14.96	14.76	16.06
# Specialist visits	7.10	11.09	7.16	11.19	6.84	11.40	7.01	11.21
Specialist>0	0.73	0.46	0.72	0.45	0.70	0.46	0.69	0.46
# Specialist visits truncated at 0	9.77	11.97	9.89	12.09	9.75	12.53	10.1	12.24
N	7,228		7,107		7,024		7,037	

fraction with at least one specialist visit both decline by about 2 percentage points, from 83% to 81% for GPs and from 72% to 70% for specialists. Individuals thus seem to avoid to contact either type of doctor after the reform. There is also weak evidence that the pattern of use might change after the reform. There are slightly more GP, and slightly less specialist visits in 2004 compared to 2003.

Figure 3.1: Development of GP and Specialist Visits Over Time



As the copayments can only be avoided by seeing neither type of physician, we are also interested in the different combinations of the outcome variables, i.e. whether an individual visited both types of doctors at least once, or only one of the two, or neither. Figure 3.1 shows the development over time of the fractions of individuals who had different combination of doctor visits in the different years. While in 2002 and 2003, roughly 10% of the sample visit neither a GP nor a specialist, this is the case for about 12% of individuals in the years after the reform. The increase in the fraction of non-visitors is accompanied by small decreases in the other three fractions.

3.4 Results

In order to compare the results of the finite mixture bivariate probit model with results from models that just focus on average effects, and from models that analyze only one of the two heterogeneities we estimate several models in addition to the one described in section 3.2. In particular, we estimate a single equation probit model where the dependent variable combines the information on the two types of doctor visits to compare our results to others in the literature. We integrate a finite mixture component with two latent classes into this model to allow for different effects in different latent classes.

Furthermore, differential effects on the two types of doctor visits are analyzed in a bivariate probit model. This model is estimated imposing a correlation of 0 between the error terms of the two underlying equations and relaxing this assumption. Similarly, when integrating the finite mixture component in the bivariate probit context, we estimate the model imposing a zero correlation and relaxing this assumption. We do this in order to investigate whether our model improves upon the one presented by Atella et al. (2004).

3.4.1 Model Selection

Table 3.2 reports AIC and BIC information criteria for the different models. As the single equation probit models use a different dependent variable than the bivariate probit models, the information criteria cannot be compared between the upper and the lower part of the table. Within the lower and upper part, however, the information in table 3.2 can help to select the model that fits the data best. In general, the table indicates that models with two latent classes fit the data better than models that do not allow for unobserved heterogeneity. Restricting ρ to zero in the bivariate models leads to a deterioration of model fit in both, the degenerate and the finite mixture case.

Among the models in the lower part of table 3.2, the finite mixture bivariate probit model that allows for correlation between the error terms within each class shows the best fit according to the AIC and BIC information criteria.

Table 3.2: Model Selection

Model	N	LogL	df	AIC	BIC
Probit Visit	28,396	-9748.09	9	19514.18	19588.47
FMM Probit Visit	28,396	-7729.19	19	15496.39	15653.21
Bivariate Probit ($\rho = 0$)	28,396	-29567.2	18	59170.4	59318.97
Bivariate Probit (ρ unrestricted)	28,396	-28076.14	19	56190.28	56347.1
FMM Bivariate Probit ($\rho_j = 0$)	28,396	-24966.48	37	50006.96	50312.36
FMM Bivariate Probit (ρ_j unrestricted)	28,396	-24887.91	39	**49853.83**	**50175.74**

Notes: Visit combines information on GP and specialists visits into one variable that is one if either type of doctor has been visited at least once.

3.4.2 Average Effects and one-way Heterogeneities

For means of comparison we begin with reporting the average marginal effects and standard errors of the single equation probit models and the simple bivariate probit model with unrestricted ρ in table 3.3. The first two columns display average marginal effects on the probability of at least one doctor visit per year, irrespective of the physician type visited. The remainder of table 3.3 reports results with the two types of heterogeneities introduced independently of each other. Columns 3 to 6 of table 3.3 report results for two latent classes in the single equation probit. The average marginal effects and standard errors displayed in columns 7 and 8 are based on the bivariate probit model where the dependent variable is split to distinguish between the probability to visit a GP and the probability to visit a specialist.

All estimations show significant changes in the probability of seeing a doctor after the increase of the copayments at the beginning of the year 2004. Marginal effects reported in column 1 of table 3.3 indicate that the probability of at least one doctor visit per year is 3 percentage points lower in 2004 and 2005 compared to 2002.

This overall effect hides heterogeneous reform effects in both analyzed dimensions. The finite mixture probit model reveals that individuals in the two latent classes react differently to the reform. The estimate of the prior probability, π_1, of belonging to latent class 1 is 0.89%. In this larger fraction of the sample the probability of at least one doctor visit only decreases by 1.9 percentage points in 2004 and 2.6 percentage points

Table 3.3: Marginal Effects after Probit, FMM Probit and Bivariate Probit

	Probit		FMM Probit				Bivariate Probit	
			Latent Class 1		Latent Class 2			
	ME	SE	ME	SE	ME	SE	ME	SE
	Pr(Visit>0)		**Pr(Visit>0)**		**Pr(Visit>0)**		**Pr(GP>0)**	
Female	0.060***	0.006	0.039***	0.003	0.120***	0.023	0.043***	0.007
Age Splines								
Age 19-40	-0.001*	0.001	-0.001	0.0003	-0.001	0.002	0.0002	0.001
Age 40-60	0.003***	0.001	0.003***	0.001	0.0004	0.004	0.004***	0.001
Age 60-80	0.001	0.001	0.0004	0.001	-0.0003	0.004	0.001	0.001
Age > 80	-0.004	0.003	0.006	0.011	0.016**	0.008	-0.004	0.003
Year Dummies								
2003	-0.004	0.003	0.001	0.003	-0.032	0.027	-0.004	0.004
2004	-0.032***	0.004	-0.019***	0.004	-0.138***	0.027	-0.019***	0.005
2005	-0.034***	0.004	-0.026***	0.004	-0.099***	0.028	-0.019***	0.005
							Pr(Specialist>0)	
Female							0.159***	0.008
Age Splines								
Age 19-40							-0.001*	0.001
Age 40-60							0.006***	0.001
Age 60-80							-0.006***	0.001
Age > 80							-0.011***	0.003
Year Dummies								
2003							-0.003	0.006
2004							-0.024***	0.006
2005							-0.030***	0.006
ρ							0.572	0.011
π_1			0.887					

* p<0.10, ** p<0.05, *** p<0.01

Notes: Marginal Effects on the probability stated at the top of each column. Visit combines information on GP and specialists visits. π_1 is the probability of class membership in latent class 1. Standard errors for the marginal effects are calculated using the delta method.

in 2005. In the other latent class the reaction amounts to 13.8 and 9.9 percentage point changes in the different years. There, thus, seems to be a rather small group of individuals who reacted sharply to the reform while the rest only changed behavior moderately.

The probabilities of different types of doctor visits are also affected differently by the reform. The results in the last columns of table 3.3 show that the probability of

seeing a specialist at least once after the reform decreases more than the probability of seeing a GP. While the probability of seeing a GP at least once decreases by 1.9 percentage points, the probability of seeing a specialist at least once is reduced by 2.4 to 3 percentage points.

3.4.3 Simultaneous Evaluation of Heterogeneities

The finite mixture bivariate probit model allows to simultaneously analyze the two types of heterogeneities. Table 3.4 presents coefficients, average marginal effects, and standard errors for the two dependent variables and the two latent classes. The estimated prior probability of belonging to latent class 1, π_1, is about 74%. There are differences in almost all parameter estimates between the two latent classes.

The sharpest difference between the two latent classes arises when focusing on the changes over time. Probabilities in all other years are again compared to the year 2002 as reference. In latent class 1 only the probability of at least one visit to a specialist in 2005 decreases significantly compared to 2002, all other year parameter estimates are statistically not different from 0. The increase of the copayments in 2004 did, thus, not affect the probability to see GPs and only had a delayed effect on the probability to see a specialist for this large fraction of the sample.

In latent class 2, on the contrary, the increase in the copayments in 2004 has changed the probability of use of both types of physicians considerably. While in 2003, individuals are as likely to visit a GP at least once as in 2002, they are 8 percentage points less likely to visit a GP in 2004. In 2005, they are still 5.5 percentage points less likely to visit a GP compared to 2002. Similarly, the probability of at least one specialist visit is reduced by 8 percentage points in 2004 and by 6 percentage points in 2005 compared to 2002. Combining the two types of heterogeneities in one model thus allows the conclusion that a large group of individuals mainly reacted through changing their probability of seeing specialists, whereas among individuals in latent class 2 the probability of both types of visits decreased significantly.

As the copayments can only be avoided entirely by not seeing any doctor, the effect of the reform on the joint probability of neither seeing a GP nor a specialist is of particular interest. Table 3.5 reports average marginal effects on the different joint probabilities for the three years 2003, 2004, and 2005 compared to 2002. Column 1 and 2 of table 3.5 display the effects and their standard errors for latent class 1, columns 3 and 4 for latent class 2, and the last two columns display the overall effect as a weighted average of the marginal effects in the two latent classes where the weight for each class is the probability of being in that class.

Table 3.4: FMM Bivariate Probit – Coefficients and Marginal Effects

	Latent Class 1				Latent Class 2			
	Coef	SE	ME	SE	Coef	SE	ME	SE
GP								
Female	0.002	0.041	0.0002	0.004	0.342***	0.046	-0.256***	0.015
Age Splines								
Age 19-40	0.011***	0.003	0.001***	0.0003	-0.004	0.005	-0.002	0.002
Age 40-60	0.018***	0.006	0.002***	0.001	0.015*	0.008	0.005*	0.003
Age 60-80	0.018	0.011	0.002	0.001	0.023***	0.008	0.008***	0.003
Age > 80	-0.114*	0.062	-0.011*	0.006	-0.009	0.020	-0.003	0.007
Year Dummies								
2003	0.070	0.049	0.006	0.004	-0.071	0.045	-0.026	0.017
2004	-0.008	0.050	-0.001	0.005	-0.214***	0.047	-0.079***	0.017
2005	-0.080	0.049	-0.008	0.005	-0.149***	0.047	-0.055***	0.018
Constant	1.129***	0.061			-0.421***	0.094		
Specialist								
Female	0.748***	0.031	0.551***	0.008	0.286***	0.041	0.103***	0.015
Age Splines								
Age 19-40	-0.004	0.003	-0.001	0.001	0.003	0.005	0.001	0.002
Age 40-60	0.027***	0.005	0.006***	0.001	0.000	0.007	0.0001	0.003
Age 60-80	-0.021***	0.006	-0.005***	0.001	-0.004	0.007	-0.001	0.002
Age > 80	-0.041***	0.012	-0.009***	0.003	-0.032	0.024	-0.012	0.009
Year Dummies								
2003	0.036	0.033	0.008	0.007	-0.077*	0.046	-0.029*	0.017
2004	-0.025	0.033	-0.005	0.007	-0.222***	0.048	-0.08***	0.017
2005	-0.092***	0.033	-0.021***	0.007	-0.172***	0.048	-0.063***	0.017
Constant	0.544***	0.052			-0.539***	0.090		
ρ	0.324***	0.026			0.069***	0.025		
π_1	0.736							

* p<0.10, ** p<0.05, *** p<0.01

Notes: π_1 is the probability of class membership in latent class 1. Standard errors for the marginal effects are calculated using the delta method.

The strongest reaction to the increase in copayments is again observed in latent class 2. The probability of no doctor visit within a year increases by 9.5 percentage points in 2004 compared to 2002 while the probability to visit both types of physicians, a specialist and a GP, at least once within the year, decreases by 6.3 percentage points. In 2005, the effects are slightly smaller, but still highly significant. In latent class 1, on the contrary, the marginal probabilities do not change significantly in 2004. In

Table 3.5: FMM Bivariate Probit – Marginal Effects on Joint Probabilities

	Latent Class 1		Latent Class 2		Overall	
	ME	SE	ME	SE	ME	SE
2003						
Pr(GP> 0, Specialist>0)	0.011	0.007	-0.023**	0.01	0.002	0.006
Pr(GP> 0, Specialist=0)	-0.005	0.006	-0.003	0.013	-0.004	0.006
Pr(GP= 0, Specialist>0)	-0.004	0.003	-0.005	0.011	-0.004	0.003
Pr(GP= 0, Specialist=0)	-0.003*	0.002	0.032**	0.014	0.006*	0.004
2004						
Pr(GP> 0, Specialist>0)	-0.005	0.008	-0.063***	0.01	-0.021***	0.006
Pr(GP> 0, Specialist=0)	0.005	0.007	-0.016	0.013	-0.001	0.006
Pr(GP= 0, Specialist>0)	0.0001	0.003	-0.017	0.011	-0.004	0.004
Pr(GP= 0, Specialist=0)	0.001	0.002	0.095***	0.015	0.026***	0.004
2005						
Pr(GP> 0, Specialist>0)	-0.024***	0.008	-0.048***	0.011	-0.031***	0.006
Pr(GP> 0, Specialist=0)	0.016**	0.007	-0.007	0.013	0.01*	0.006
Pr(GP= 0, Specialist>0)	0.004	0.003	-0.014	0.011	-0.001	0.004
Pr(GP= 0, Specialist=0)	0.005**	0.002	0.069***	0.015	0.022***	0.004

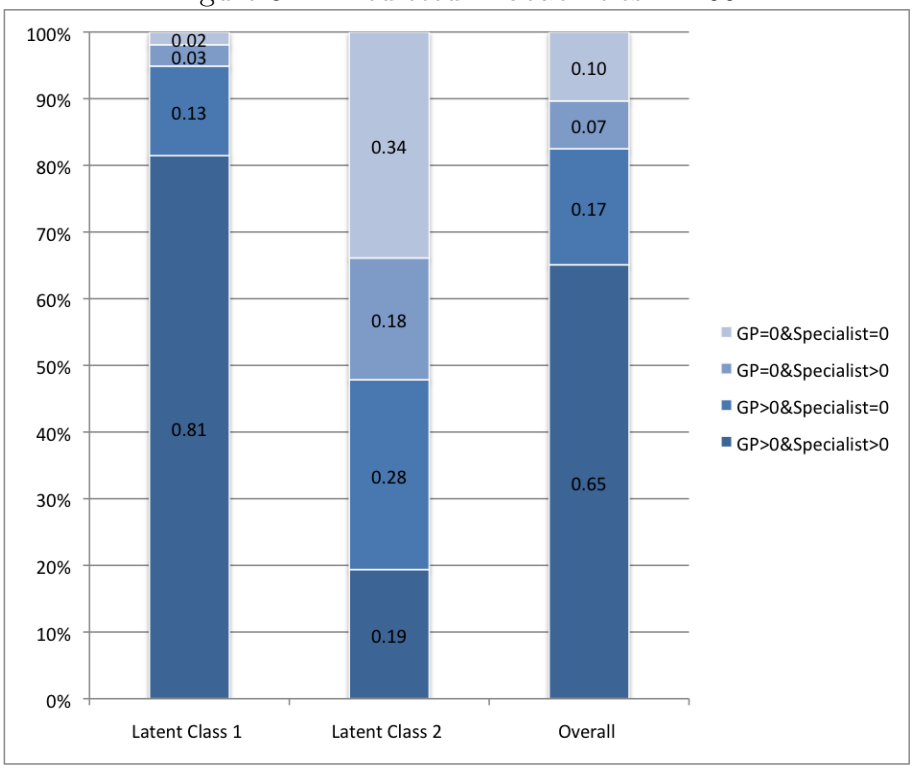
* p<0.10, ** p<0.05, *** p<0.01

Notes: Marginal changes are calculated with the respective probabilities in 2002 as reference. Standard errors for the marginal effects are calculated using the delta method.

2005, the probability of seeing a GP and a specialist is reduced by 2.4 percentage points compared to 2002 while at the same time the probability of seeing only GPs increases by 1.6 percentage points and the probability of no doctor visit increases by .5 percentage points.

The marginal effects on the joint probabilities, thus, reveal further differences between the two latent classes. In latent class 2 individuals immediately react to the increase in the copayments by reducing their probability to see any doctor. In latent class 1 the reaction occurs only in 2005, and occurs mainly through focussing on only one type of visit, in particular on GP visits. This reaction could be explained by the referral to other doctors that is needed in order to avoid the per-quarter fee. When an individual has already paid the fee at the first visit in a quarter, additional visits to other doctors are only exempt from the fee if the already consulted doctor writes a referral for the specific other type of physician. The question whether the delayed reaction in latent class 1 could be due to changes in the practice of writing referrals in 2005 is left for further research.

Figure 3.2: Predicted Probabilities in 2002



The results suggest that the introduction of the per-quarter fee might have strengthened the role of the GP in the German statutory health insurance. While GPs are not officially implemented as gatekeepers, i.e. individuals can choose to directly visit a specialist for a given disease and this would not cost them more than visiting a GP first, the reform nevertheless induces a large part of the sample to focus on GP visits. This might highlight the possibility of introducing gatekeeping into a health insurance system in a way that does not limit the patients' choices.

Naturally, the question arises who the individuals are that belong to the different latent classes. Based on our model's predictions, we can compare the two latent classes in terms of the probability of use of doctor visits. Figure 3.2 displays the predicted probabilities of using the different types of physicians in 2002. It shows that in latent class 1 the probability of no doctor visit within the year 2002 is about 2 percent. Most of these individuals thus go to a doctor at least once within a year. And most of them see both, a GP and a specialist at least once. The probability to see both types of physicians is 81 percent. In latent class 2, the picture looks very different. The predicted probability to see no physician at all within the year 2002 is 34 percent, the probability to see both types of physicians is only 19 percent. Latent class 1 might thus be described as the group of *likely users*, while individuals in latent class 2 are

the *less-likely users*.

Overall, we find evidence for heterogeneous effects of increases in copayments. When analyzing the two types of heterogeneities independently of each other, we find that a large class of *likely users* of health care reacts much less to the increase in copayments than the small class of *less-likely users*. Furthermore, the probability of at least one visit to a GP decreases less than the probability of at least one visit to a specialist. Simultaneously analyzing both types of heterogeneities reveals that the *likely users* not only react less strongly to the increase in copayments, but also by changing their pattern of use. While the *less-likely users* visit either type of physician with a smaller probability after the reform, *likely users* focus on one type of visit: Instead of seeing GPs and specialists in a year, they only tend to visit GPs.

3.5 Conclusion

The interest in heterogeneous reform effects has recently increased in the literature. In order to evaluate heterogeneous effects, more sophisticated econometric models are needed. In this chapter we develop and apply a finite mixture bivariate probit model that allows to estimate effects on two binary outcomes in different latent groups in the population.

We apply our model to estimate the effects of an increase in copayments in the German statutory health insurance in 2004. In particular, we analyze whether the reform affected access to different types of health care differently in two latent groups. The different types of access to health care are measured as the probabilities of at least one visit to a GP and of at least one visit to a specialist within a year. As earlier studies of this reform did not distinguish between different types of doctor visits, we also present results on general access to health care, measured as the probability of at least one doctor visit per year, irrespective of the type of doctor visited. Furthermore, we gauge the contribution of our model by also analyzing the effects on the two types of doctor visits without distinguishing between latent classes.

Using insurance claims data from the largest German sickness fund, we find that the probability to visit any kind of doctor at least once within a year decreases by about 3 percentage points after the increase in the copayments. Distinguishing the results between GP and specialist visits shows that the probability of at least one GP visit decreases less than the probability of at least one specialist visit. Furthermore, we can distinguish two underlying groups: *likely users* who react less to the increase in copayments and *less-likely users* who show a strong reduction in the probability

to visit a doctor. This result aligns with findings in the literature that the demand for health care is more elastic among low users of health care than among high users (Bago d'Uva, 2006; Winkelmann, 2006).

Combining the analysis of the two types of heterogeneities into one model allows to uncover that the *likely users* not only react less to the increase in copayments than the *less-likely users*. They also react in a different way: While the *less-likely users* tend to go to neither type of doctor after the reform, the *likely users* change their behavior by focussing on one type of doctor visits. They are more likely to visit only a GP after the increase of the copayments and less likely to see both, a specialist and a GP. This might indicate that per-quarter fees for doctor visits have the potential to concentrate visits to one type of doctor. In particular, they may help to establish GPs as gatekeepers without actually forcing individuals to visit a GP first.

Appendix: Marginal Effects

Marginal effects on the probabilities of each type of doctor visit and on the joint probabilities for the different combinations of the different types of visits are calculated. Standard errors for the marginal effects are derived using the delta method. For continuous explanatory variables the marginal effects are calculated using the calculus method. Marginal effects of binary variables are calculated with the finite difference method.

For a continuous variable x the marginal effect on the different joint probabilities for y_{1i} and y_{2i} in latent class j are calculated for individual i

$$\begin{aligned} ME_{xi}Pr(y_{1i}, y_{2i}|x_i)_j &= \frac{\partial \Phi_2(q_{1i}x'\beta_j, q_{2i}x_i'\gamma_j, q_{1i}q_{2i}\rho_j)}{\partial x_i} \\ &= q_{1i}\beta_{x,j}\phi(q_{1i}x_i'\beta_j)\Phi\left(\frac{q_{2i}x_i'\gamma_j - q_{1i}^2 q_{2i}\rho_j x_i'\beta_j}{\sqrt{1-\rho_j^2}}\right) \\ &+ q_{2i}\gamma_{x,j}\phi(q_{2i}x_i'\gamma_j)\Phi\left(\frac{q_{1i}x_i'\beta_j - q_{1i}q_{2i}^2\rho_j x_i'\gamma_j}{\sqrt{1-\rho_j^2}}\right) \end{aligned} \tag{3.3}$$

where Φ_2 stands for the cumulative bivariate normal distribution function, Φ and ϕ are the univariate standard normal cumulative distribution function and density respectively, and $q_{ki} = 2y_{ki} - 1$, where $k \in \{1, 2\}$. The values are averaged over all individuals to get the average marginal effect.

Marginal effects of the continuous variable x on the marginal probabilities of the outcomes in latent class j for individual i are calculated as

$$ME_{xi}Pr(y_{1i} = 1|x_i)_j = \beta_{x,j}\phi(x_i'\beta_j) \tag{3.4}$$

$$ME_{xi}Pr(y_{2i} = 1|x_i)_j = \gamma_{x,j}\phi(x_i'\gamma_j) \tag{3.5}$$

The individual marginal effects on the joint outcome probabilities of the years 2003, 2004 and 2005 compared to 2002 in latent class j for individual i are calculated as

$$\begin{aligned} ME_{year,i}Pr(y_{1i}, y_{2i}|x_i)_j = \Phi_2\left(q_{1i}(x_i'\beta_j + \beta_{year,j}), q_{2i}(x_i'\gamma_j + \gamma_{year,j}), q_{1i}q_{2i}\rho_j\right) \\ - \Phi_2(q_{1i}x_i'\beta_j, q_{2i}x_i'\gamma_j, q_{1i}q_{2i}\rho_j) \end{aligned} \tag{3.6}$$

where q_{ki} is defined as above, and β_j and γ_j are vectors of parameter estimates for all variables but the year indicators. Again the averages of these marginal effects over

all individuals are reported.

Marginal effects of the year dummies on the marginal probabilities for each individual i are derived as

$$ME_{year,i}Pr(y_{1i} = 1|x_i)_j = \Phi(q_{1i}(x_i'\beta_j + \beta_{year,j})) - \Phi(q_{1i}x_i'\beta_j) \tag{3.7}$$

$$ME_{year,i}Pr(y_{2i} = 1|x_i)_j = \Phi(q_{2i}(x_i'\gamma_j + \gamma_{year,j})) - \Phi(q_{2i}x_i'\gamma_j) \tag{3.8}$$

The overall marginal effect, i.e. the marginal effect averaged over the latent classes, for any continuous or discrete variable x is then derived as weighted average of the marginal effects across the different latent classes

$$ME_{xi} = \sum_j \pi_j ME_{xi,j} \tag{3.9}$$

Averages over all individuals are reported.

Standard errors for the average marginal effects are derived using the delta method that delivers the variance for each average marginal effect as

$$Var(ME) = \nabla_g' Var(\theta) \nabla_g \tag{3.10}$$

where θ is the vector of all parameters that are estimated (β_j,γ_j,ρ_j, and p, where $\pi = \frac{exp(p)}{1+exp(p)}$) and ∇_g stands for the gradient of the marginal effect, $ME = g(\theta)$, with respect to θ. In order to calculate the variance on the average marginal effect, we set each element in the gradient to its sample average.

Chapter 4

Is traditional teaching really all that bad? A within-student between-subject approach

4.1 Introduction

Recent studies stress the importance of teachers for student learning. However, the question what actually determines teacher quality, i.e. what makes one teacher more successful in enhancing her students' performance than another, has not been settled so far (Aaronson et al., 2007). Different categories of teacher variables have been analyzed. Some studies focus on the impact of a teacher's gender and race on teacher quality (Dee, 2005, 2007). Others try to uncover the relationship between student outcomes and teacher qualifications such as teaching certificates, other paper qualifications or teaching experience (Kane et al., 2008). Such observable teacher characteristics are, however, generally found to have only little impact on student achievement and can only explain a relatively small part of overall teacher quality (Aaronson et al., 2007; Rivkin et al., 2005). Most of the variation in teacher quality can be attributed to unobserved factors.[1]

While most of these studies focus on characteristics of the teacher, this chapter directly peers into the black box of educational production by focusing on the actual teaching process. More specifically, we contrast two teaching practices (as currently implemented in practice), teaching by giving lecture style presentations with teaching based on in-class problem solving, and investigate the impact on student achievement.

[1]This finding led researchers, concerned with providing recommendations for recruitment policies and designing optimal teacher pay schemes, to suggest to identify effective teachers by their actual performance on the job using "value added" measures of student achievement (Gordon et al., 2006).

Giving lecture style presentations is often regarded as old-fashioned and connected with many disadvantages: Lectures fail to provide instructors with feedback about student learning and rest on the presumption that all students learn at the same pace. Moreover, students' attention wanes quickly during lectures and information tends to be forgotten quickly when students are passive. Finally, lectures emphasize learning by listening, which is a disadvantage for students who prefer other learning styles. Alternative instructional practices based on active and problem-oriented learning presumably do not suffer from these disadvantages. National standards (NCTM, 1991; National Research Council, 1996) consequently advocate engaging students more in hands-on learning activities and group work. Despite these recommendations traditional lecture and textbook methodologies continue to dominate science and mathematics instruction in US middle schools (Weiss, 1997). This raises the question whether student achievement could be raised by reducing the high share of teaching time devoted to lecture style presentations.

By addressing this question, this study adds to the literature analyzing the impact of teaching process variables such as teaching practices on student outcomes.[2] Despite the importance of teaching practices for student performance as recognized by educational researchers (Seidel and Shavelson, 2007) and their potential relatively low-cost implementation, economists have only recently begun to analyze the impact of teaching methods on student achievement.[3] Various dimensions of teaching practices have been shown to be able to explain a large share of the between-teacher variation in student achievement (Schacter and Thum, 2004). However, to our knowledge no rigorous empirical analysis of the effect of lecture style teaching compared to in-class problem solving as implemented in practice today exists.[4]

To study the effect of lecture style teaching relative to teaching based on in-class problem solving we use information on in-class time use provided by teachers in the 2003 wave of the Trends in International Mathematics and Science Study (TIMSS) in US schools. Estimating a reduced form educational production function and exploiting between-subject variation to control for unobserved student traits, we find that the choice of teaching practices matters for student achievement. We find that a 10

[2]For an overview see Goe (2007).

[3]Brewer and Goldhaber (1997) and Aslam and Kingdon (2007) analyze the impact of many different teaching methods. Rouse and Krueger (2004), Banerjee et al. (2007), and Barrow et al. (2009) investigate the effectiveness of computer-aided instruction and Machin and McNally (2008) analyze an education policy that changed reading instruction.

[4]Note that our results do not allow any comparison between teaching practices when implemented ideally, but only allow a comparison of lecture-style teaching with problem-style learning approaches as implemented in practice.

percentage point shift from problem solving to lecture style presentation results in an increase in student achievement of about 1 percent of a standard deviation.

This result is highly robust. Consistent with other studies in this literature, we find no evidence for significant effects of commonly investigated observable teacher characteristics such as teaching certificates or teaching experience. While we are able to control for a huge array of observable teacher traits, selection of teachers based on unobservable characteristics into teaching methods remains an issue. The bias resulting from potential selection of teachers with different unobservable attributes into different teaching methods is assessed following the technique pioneered in Altonji et al. (2005). The results indicate that only relatively low selection on unobservables compared to the selection on observables is necessary to explain the entire estimated effect. We would thus not formulate policy conclusions that call for more lecture style teaching compared to problem solving in general. However, a negative causal effect of giving lecture style presentations that is hidden in our results due to selection based on unobserved teacher traits is also not very likely. It can only exist if "good" teachers (teachers with favorable unobserved characteristics) predominately select themselves into an inferior teaching technique. This scenario, however, lacks any intuitive or theoretical support and thus appears extremely implausible. We therefore do not find any evidence that the high share of total teaching time devoted to traditional lecture style teaching in science and mathematics instruction in US middle schools has detrimental effects on student achievement. Our findings imply that simply changing the teaching method from lecture style presentation to problem solving without concern for how the methods are implemented has little potential for raising overall achievement levels.

The remainder of the chapter is structured as follows: the following section reviews the literature on teaching practices. Section 3 presents the data. Section 4 describes the estimation strategy. Headline results are presented in Section 5, while Section 6 provides a sensitivity analysis. Section 7 concludes.

4.2 Literature on Teaching Practices

There are two strands of literature that are closely related to our study. The first strand analyzes the impact of different teaching styles on student achievement. In these studies teaching style variables are meant to proxy for broader pedagogical concepts. A certain teaching style may consist of a combination of different teaching practices, where the term teaching practice indicates the actual classroom activity, like giving a lecture style presentation or reviewing homework. The effects of single teaching practices are

analyzed by the second strand of the related literature.

Findings of the first strand of literature on teaching style speak in favor of modern, interactive teaching styles. Slavin et al. (2009) review recent research on the achievement outcomes of mathematics programs for middle and high schools. The most striking observation of this review is the evidence supporting instructional process strategies, especially cooperative learning. Several studies reveal important impacts of cooperative learning programs that target teachers' instructional behaviors rather than math content alone. The review concludes that educators and researchers might do well to focus more on what is actually happening in the classroom.

In an earlier meta-analysis, Lou et al. (1996) focus on a typical component of cooperative learning strategies, namely within-class grouping. The authors conclude that within-class grouping has the potential to raise student achievement. In particular, it works best when the physical placement of students into groups is combined with the adaptation of instruction methods and materials for small-group learning. Dignath and Büttner (2008) find negative impacts of group work for primary school students in another meta-analysis. While these results are somehow in contradiction to the findings by Lou et al. (1996) and Slavin et al. (2009), the authors emphasize that presumably the implementation of cooperative learning strategies into real classroom settings matters.

Results from a study by Machin and McNally (2008) also support the idea that coordinated and comprehensive changes in instructional practices are beneficial for student learning. This study analyzes the effects of the introduction of the "literacy hour" in English primary schools in the late 1990s. This policy intervention by the British government changed how primary school students are taught to read. Using the fact that not all schools started the literacy hour at the same point in time in a difference-in-difference framework, the authors show that the literacy hour significantly increased reading skills for low ability student while high ability students were not affected.

The findings of these studies suggest that the overall teaching style matters for student achievement. In particular, modern, interactive teaching styles have the potential to positively affect student learning. However, the implementation of more interactive teaching strategies into real classroom settings matters. There appear to be important interaction effects between individual instruction practices, curricula, teaching materials, and teacher experience with more interactive teaching styles.

The second strand of literature analyzes individual teaching practices more specifically. A first group of studies in this area analyzes the use of computer based instruc-

tion. The evidence on the use of computer programs is mixed for different subjects. On the one hand, Rouse and Krueger (2004) find that literacy skills of 3rd and 6th grade students in a large US urban school district do not profit from the use of computer programs for teaching language and reading. Students' math skills, on the other hand, are found to increase due to computer-aided instruction as shown by Banerjee et al. (2007) for India and Barrow et al. (2009) for the US. Computer-based instruction might thus be more effective in some than in other subjects.

Other teaching practices that have received some attention in the literature are accountability measures. Wenglinsky (2002) finds that frequent testing of students is positively related to students' test scores taking into account student background and prior performance. Bonesrønning (2004) analyzes if grading practices affect student achievement in Norway and finds evidence that easy grading deteriorates student achievement. The design of tests might also matter. This hypothesis is supported by findings of Matsumura et al. (2002). This study finds evidence for a positive link between the quality of assignments that students are asked to do and overall student performance. These findings emphasize the importance of taking into account other categories of in-class time use such as accountability measures when studying the effect of teaching by giving lecture style presentations in comparison to in-class problem solving.

More closely related to this chapter in terms of the teaching practices analyzed and identification strategy are the analyses by Brewer and Goldhaber (1997) and Aslam and Kingdon (2007). Brewer and Goldhaber (1997) estimate different specifications of education production functions for tenth grade students in math with data from the National Educational Longitudinal Study of 1988. They conclude that teacher behavior is important in explaining student test scores. Controlling for student background, prior performance and school and teacher characteristics, they find that instruction in small groups and emphasis on problem solving lead to lower student test scores.

Aslam and Kingdon (2007) analyze the impact of different teaching practices on student achievement in Pakistan. Their identification strategy rests on within student across subject (rather than across time) variation, which is similar to the identification strategy employed in this analysis. They find that students taught by teachers who spend more time on lesson planning and by teachers who ask more questions in class have higher test scores.

Similar to this second strand of studies we focus on a comparison of single teaching practices, namely time devoted to teaching by giving lecture style presentations versus time devoted to teaching by problem solving. As lecture style teaching is a teaching

practice that is often associated with more traditional, didactic or teacher-centered teaching styles, while problem solving is connected to more modern, interactive or student-centered approaches to teaching, our results might be also of a more general interest. Strictly speaking, however, our data only allows to draw conclusions about actual classroom activities.

4.3 Data

The data used in this study is the 2003 wave of the Trends in International Mathematics and Science Study (TIMSS). In this study we focus on country information for the US. In TIMSS, students in 4th grade and in 8th grade were tested in math and science. We limit our analysis to 8th grade students, because 4th grade students are typically taught by one teacher in all subjects.

We standardize the test scores for each subject to be mean 0 and standard deviation 1. In addition to test scores in the two tested subjects, the TIMSS data provide background information on student home and family. For the purpose of this analysis it is crucial that TIMSS allows linking students to teachers. Each student's teachers in math and science are surveyed on their characteristics, qualifications and teaching practices. Additionally, school principals provide information on school characteristics.

The key variable of interest in this chapter is derived from question 20 in the teacher questionnaires in the 2003 wave of TIMSS. Unfortunately, this precise question was not asked in previous waves of TIMSS. We therefore limit our analysis to the 2003 wave. Teachers are asked in 2003 to report what percentage of time in a typical week of the specific subject's lessons students spend on eight in-class activities. Our three main categories of interest are listening to lecture style presentation, working on problems with the teacher's guidance and working on problems without guidance. The overall time in class spent on these three activities likely provides a good proxy for the time in class in which students are taught new material. The main interest of this study is to contrast teaching by giving lecture style presentations to teaching based on problem solving. The two problem solving categories are thus combined into a single teaching practice: teaching based on problem solving. To allow a comparison of the two categories we construct a single variable, *lecture style teaching*, that is the percent of time spent giving lecture style presentation relative to the percent of time spent on either of the two activities ($\frac{lecture(\%)}{lecture(\%)+problemsolving(\%)}$). The key advantage is that a change in this variable can directly be interpreted as a shift from spending time on one to spending time on the other practice holding constant the overall time spent

on these two practices. For example, an increase in this variable of 0.1 indicates that 10 percentage points of total time devoted to teaching new material are shifted from teaching based on problem solving to giving lecture style presentations.

The remaining 5 activities include reviewing homework, listening to the teacher reteach and clarify content, taking tests or quizzes, classroom management and other activities. We simply group these activities and construct a single variable, *other class activities*, which measures the share of total time in class spent on other activities. We include this variable as a control in all specifications. To analyze the robustness of our results we also estimate specifications that include the individual shares of all categories. Moreover, teachers also report the total time in minutes per week that they teach math or science to the class, which we also include as a control variable in most specifications.

The TIMSS 2003 US data set contains student-teacher observations on 8,912 students in 232 schools. 41 of those students have more than one teacher in science. These students are not included in the estimation sample. 8,871 students in 231 schools in 455 math classes taught by 375 different math teachers and in 1,085 science classes taught by 475 different science teachers remain in the sample. Not all of the students and teachers completed their questionnaires. In order not to lose a large amount of observations we impute missing values of all control variables and include indicators for imputed values in all estimations.[5] 2,561 students have, however, missing information on our teaching practice variable of interest. These observation are dropped from the analysis. 6,310 students in 205 schools with 639 teachers (303 math teachers and 355 science teachers, where 19 teachers teach both subjects) remain in the sample.

Furthermore due to the sampling design of TIMSS, students are not all selected with the same probability. A two stage sampling design makes it necessary to take probability weights into account when estimating summary statistics (Martin, 2005). All estimation results take the probability weights into account and allow for correlation between error terms within schools.[6]

Table 4.1 reports descriptive statistics on observable teacher characteristics separately for math and science teachers. Mean differences are reported in the last column of table 4.1. The share of teachers with math or science majors naturally differs be-

[5]Experimenting with different imputation procedures revealed that our main results do not depend on the method of imputation. Main results remain also qualitatively unchanged when simply deleting observations with missing values. Results presented in the chapter are based on a simple mean-imputation procedure.

[6]In addition, the two step procedure of sampling could be incorporated in the estimation of standard errors. For simplicity, we ignore the latter in the following analysis which then gives us conservative estimates of the standard errors.

Table 4.1: Descriptive Statistics – Teacher variables

	Math		Science		
	303 teachers		355 teachers		
Variable	Mean	SD	Mean	SD	Difference
In-class time use					
Lecture style teaching	0.320	0.187	0.374	0.202	-0.054***
Other class activities	0.428	0.119	0.446	0.161	-0.018
Total teaching time ($\frac{min}{week}$)	226.27	43.89	223.72	47.34	2.55
Teacher variables					
Teacher is female	0.649	0.473	0.540	0.496	0.109**
Full teaching certificate	0.970	0.163	0.957	0.188	0.013
Major in math	0.473	0.492	0.099	0.294	0.374***
Major in science	0.146	0.348	0.584	0.486	-0.438***
Major in education	0.598	0.483	0.456	0.491	0.142***
Teacher younger than 30	0.119	0.322	0.143	0.349	-0.024
Teacher aged 30-39	0.273	0.443	0.223	0.414	0.050
Teacher aged 40-49	0.293	0.452	0.335	0.469	-0.042
Teaching experience $<$ 1 year	0.043	0.201	0.042	0.199	0.001
Teaching experience 1-5 years	0.178	0.370	0.224	0.404	-0.046
Teacher training 0 years	0.102	0.301	0.154	0.359	-0.051**
Teacher training 1 year	0.578	0.491	0.523	0.497	0.055
Teacher training 2 years	0.209	0.404	0.193	0.392	0.016
Teacher training 3 years	0.039	0.192	0.048	0.213	-0.009
Teacher training 4 years	0.056	0.228	0.035	0.184	0.021
Teacher training 5 years	0.008	0.090	0.039	0.193	-0.031**
Motivation					
Pedagogy classes in last 2 years	0.748	0.431	0.648	0.472	0.100***
Subject content classes in last 2 years	0.840	0.364	0.827	0.374	0.014
Subject curriculum classes in last 2 years	0.830	0.372	0.853	0.349	-0.023
Subject related IT classes in last 2 years	0.729	0.441	0.803	0.393	-0.074**
Subject assessment classes in last 2 years	0.756	0.426	0.649	0.471	0.107***
Classes on improving student's skills last 2 years	0.759	0.424	0.766	0.418	-0.007
Working hours scheduled per week	21.12	8.28	20.16	7.29	0.960
Weekly hours spent on lesson planning	3.70	2.71	4.68	3.28	-0.976***
Weekly hours spent on grading	5.25	3.93	6.08	4.41	-0.830**
Teaching requirements					
Requirement probationary period	0.502	0.493	0.496	0.479	0.007
Requirement licensing exam	0.526	0.493	0.558	0.479	-0.032
Requirement finished Isced5a	0.891	0.307	0.824	0.371	0.067**
Requirement minimum education classes	0.833	0.368	0.777	0.399	0.056
Requirement minimum subject specific classes	0.799	0.395	0.744	0.420	0.056

Note: Probability weights and within school correlation are taken into account when estimating means and standard deviations. Teacher variables are weighted by the number of students taught by each teacher.

tween those two groups. Apart from mean differences in majors several other variables are significantly different between the two groups. We control for all these observable differences in teacher traits in our empirical analysis.

To investigate which teachers teach using relatively more lecture style presentations and which students are more exposed to this method tables 4.2 and 4.3 display averages of student, school, class and teacher characteristics grouped by the share of lecture style presentation. The last column of tables 4.2 and 4.3 presents the mean differences between characteristics of students, schools, classes and teachers with more and less than the median share of time spent on lecture style presentation.

Table 4.2 reveals that the only student characteristic which shows significant differences between the two groups is the number of books that families have at home. Students who are taught with relatively less time spent on lecture style presentations seem to have fewer books at home. Furthermore, the share of students in schools for which the principal reports very low involvement of parents in school activities is higher among students taught with relatively less time spent on lecture style presentations. Students exposed to smaller shares of class time devoted to lecture style teaching are also slightly more likely to be in tracked classes.

Table 4.3 reports teacher characteristics by the intensity of lecture style teaching. Students taught with relatively less time spent on lecture style teaching have a higher share of female teachers, more teachers who are at least 50, teachers who have the maximum number of years of teacher training, and teachers who have taken pedagogical or content knowledge classes in the last two years. All other teacher characteristics do not differ significantly between the two groups.

Tables 4.2 and 4.3 indicate some potential for selection into different teaching practices. To control for the confounding impact of these differences in observable characteristics, we include all variables presented in the tables above and additional information on teaching limits (see table 4.8 in the appendix) in our empirical analysis. The latter are teacher reports on limitations that they face in teaching. These include limitations through the class composition, shortages in teaching material, and a high student teacher ratio.

4.4 Estimation Strategy

To estimate the effect of giving lecture style presentations relative to teaching based on problem solving we estimate a standard education production function:

Table 4.2: Student, School and Class Variables by Intensity of Lecture Style Teaching

	Lecture Style > median		**Lecture Style <= median**		
Variable	Mean	SD	Mean	SD	Difference
Student					
Age	14.22	0.451	14.23	0.470	-0.014
Born in first 6 months	0.501	0.499	0.495	0.499	0.005
Girl	0.522	0.500	0.522	0.499	0.000
Number books at home: 11-25	0.175	0.378	0.173	0.376	0.002
Number books at home: 26-100	0.282	0.448	0.278	0.445	0.004
Number books at home: 101-200	0.185	0.387	0.173	0.375	0.013*
Number books at home: >200	0.248	0.430	0.250	0.430	-0.002
Parental Education lower secondary	0.057	0.208	0.058	0.211	-0.001
Parental Education upper secondary	0.248	0.388	0.253	0.392	-0.005
Parental Education post secondary voc/technical	0.094	0.264	0.088	0.254	0.006
Parental Education university	0.577	0.445	0.571	0.446	0.006
Number people at home 3	0.165	0.368	0.170	0.372	-0.005
Number people at home 4	0.347	0.472	0.350	0.472	-0.003
Number people at home 5	0.255	0.432	0.251	0.429	0.004
Number people at home 6	0.108	0.307	0.104	0.303	0.003
Number people at home 7	0.046	0.207	0.042	0.198	0.004
Number people at home 8 or more	0.039	0.192	0.039	0.193	0.000
Never speaks English at home	0.044	0.205	0.048	0.213	-0.004
Sometime speaks English at home	0.099	0.298	0.093	0.289	0.006
Almost always speaks English at home	0.845	0.360	0.843	0.361	0.002
Immigrant	0.141	0.347	0.144	0.349	-0.003
School					
Community, 3001-15,000 people	0.188	0.375	0.228	0.403	-0.041
Community, 15,001-50,000 people	0.330	0.454	0.293	0.435	0.037
Community, 50,001 - 100,000 people	0.121	0.314	0.118	0.309	0.003
Community, 100,001 - 500,000 people	0.137	0.331	0.130	0.321	0.007
Community, more than 500,000 people	0.125	0.319	0.111	0.299	0.014
Parental involvement in school - very low	0.130	0.319	0.204	0.384	-0.074**
Parental involvement in school - low	0.389	0.469	0.324	0.442	0.065
Parental involvement in school - medium	0.300	0.440	0.316	0.441	-0.016
Parental involvement in school - high	0.118	0.311	0.101	0.285	0.017
Class variables					
Total teaching time ($\frac{min}{week}$)	221.93	48.60	227.45	43.03	-5.52
Class size	23.95	6.90	24.03	7.20	-0.073
Tracked according to ability	0.319	0.452	0.394	0.469	-0.075*

Note: Lecture style measures the share of teaching time devoted to lecture style teaching instead of teaching based on problem solving. Median refers to the median of lecture style teaching. Probability weights and within school correlation are taken into account when estimating means and standard deviations.

Table 4.3: Teacher Variables by Intensity of Lecture Style Teaching

	Lecture Style > median		**Lecture Style <= median**		
Variable	Mean	SD	Mean	SD	Difference
Teacher variables					
Teacher is female	0.535	0.496	0.642	0.475	-0.107**
Full teaching certificate	0.947	0.216	0.977	0.135	-0.031
Major in math	0.265	0.432	0.303	0.455	-0.039
Major in science	0.385	0.477	0.349	0.471	0.036
Major in education	0.515	0.490	0.537	0.493	-0.022
Teacher younger than 30	0.127	0.331	0.134	0.339	-0.007
Teacher aged 40-49	0.367	0.480	0.272	0.442	0.096**
Teacher at least 50	0.256	0.434	0.347	0.473	-0.091**
Teaching experience < 1 year	0.045	0.205	0.041	0.197	0.004
Teaching experience 1-5 years	0.229	0.406	0.180	0.372	0.049
Teacher training 0 years	0.130	0.334	0.127	0.331	0.003
Teacher training 1 year	0.568	0.492	0.537	0.496	0.031
Teacher training 2 years	0.213	0.407	0.191	0.391	0.022
Teacher training 3 years	0.036	0.185	0.050	0.216	-0.013
Teacher training 4 years	0.040	0.195	0.050	0.216	-0.009
Teacher training 5 years	0.008	0.089	0.036	0.186	-0.028**
Teacher motivation					
Pedagogy classes in last 2 years	0.633	0.476	0.749	0.429	-0.116***
Subject content classes in last 2 years	0.792	0.402	0.865	0.338	-0.073**
Subject curriculum classes in last 2 years	0.818	0.381	0.860	0.342	-0.042
Subject related IT classes in last 2 years	0.780	0.409	0.754	0.427	0.025
Subject assessment classes in last 2 years	0.671	0.464	0.728	0.441	-0.057
Classes on improving student's skills last 2 years	0.757	0.424	0.768	0.418	-0.011
Working hours scheduled per week	20.82	7.45	20.50	8.09	0.322
Weekly hours spent on lesson planning	4.29	3.14	4.12	2.96	0.170
Weekly hours spent on grading	5.66	3.85	5.67	4.45	-0.013
Teaching requirements					
Requirement probationary period	0.503	0.484	0.496	0.487	0.008
Requirement licensing exam	0.523	0.486	0.556	0.487	-0.033
Requirement finished ISCED 5a	0.858	0.340	0.857	0.343	0.001
Requirement minimum education classes	0.779	0.402	0.826	0.369	-0.047
Requirement minimum subject specific classes	0.740	0.425	0.797	0.392	-0.056

Note: Lecture style measures the share of teaching time devoted to lecture style teaching instead of teaching based on problem solving. Median refers to the median of lecture style teaching. Probability weights and within school correlation are taken into account when estimating means and standard deviations. Teacher variables are weighted by the number of students taught by each teacher.

$$Y_{ijk} = c_j + B'_{ijk}\beta_{1j} + S'_{ijk}\beta_{2j} + T'_{ijk}\beta_{3j} + Lecture_{ijk}\beta_{4j} + \epsilon_{ijk}. \tag{4.1}$$

The test score, Y_{ijk}, of student i in subject j in school k is explained by a subject specific constant c_j, student background characteristics, B_{ijk}, school characteristics, S_{ijk}, teacher and class characteristics, T_{ijk}, and the variable $Lecture_{ijk}$. The last variable constitutes the focus of this analysis. It represents teaching time spent on lecture style presentation relative to problem solving. The error term, ϵ_{ijk}, contains all unobservable influences on student test scores. In particular, it contains the effects of unobservable student, μ_i , teacher, ξ_j, and school characteristics, ν_k:

$$\epsilon_{ijk} = \mu_i + \xi_j + \nu_k + \psi_{ijk} \tag{4.2}$$

Estimating equation (4.1) by ordinary least squares produces biased estimates if unobserved school characteristics, ν_k, and $Lecture_{ijk}$ are correlated. This can be the case if the choice of the teaching practice is partly determined by the school and if there exists sorting of high ability students or effective teachers into schools.

To eliminate the effects of between-school sorting, we use school fixed effects, s_k, to exclude any systematic between-school variation in performance or teaching practice, whatever its source:

$$Y_{ijk} = c_j + B'_{ik}\beta_{1j} + s_k + T'_{ijk}\beta_{3j} + Lecture_{ijk}\beta_{4j} + \mu_i + \xi_j + \psi_{ijk}. \tag{4.3}$$

The estimates produced by equation (4.3) could still be biased by within-school sorting wherever schools have more than one class per subject per grade. We therefore eliminate the influence on constant student traits by differencing between subjects:

$$\begin{aligned} \Delta Y_i \;=\; & c_m - c_s + B'_i(\beta_{1m} - \beta_{1s}) + S'_i(\beta_{2m} - \beta_{2s}) \\ & + T'_{im}\beta_{3m} - T'_{is}\beta_{3s} + Lecture_{im}\beta_{4m} - Lecture_{is}\beta_{4s} + \eta_i \end{aligned} \tag{4.4}$$

where $\Delta Y_i = Y_{im} - Y_{is}$ and $\eta_i = \xi_m - \xi_s + \psi_{im} - \psi_{is}$.

In our headline specification we follow Dee (2005, 2007) by assuming that coefficients for each variable are equal across the two subjects:[7]

$$\Delta Y_i \;= \Delta T'_i\beta_3 + \Delta Lecture_i\beta_4 + \eta_i. \tag{4.5}$$

[7]We do, however, estimate equation (4.4) as a robustness check.

The estimate of the effect of teaching practice on student achievement produced by equation (4.5) is not biased due to between or within school sorting of students based on unobservable student traits. We do, however, have to make the identifying assumption that unobservable teacher characteristics that directly influence student achievement are not related to the choice of the teaching method. In other words, η_i is uncorrelated with all other right-hand side variables. This is a strong assumption and we therefore refrain from interpreting β_4 as causal effect. We rather interpret β_4 as a measure for the link between a teaching practice and student achievement that is not driven by between or within school sorting of students. It might, however, be partly driven by sorting of teachers into a special teaching method based on unobservable teacher traits.

We evaluate the concern of selection on unobservables by borrowing a procedure from Altonji et al. (2005) which allows to evaluate the bias of the estimate under the assumption that selection on unobservables occurs to the same degree as selection on observables. As developed in the appendix the asymptotic bias of $\widehat{\beta_4}$ in our application is

$$Bias(\widehat{\beta_4}) = \frac{Cov(\Delta\widetilde{Lecture}, \eta)}{Var(\Delta\widetilde{Lecture})} = \frac{Cov(\Delta Lecture, \eta)}{Var(\Delta\widetilde{Lecture})} \tag{4.6}$$

where indices are omitted for simplicity and where $\Delta\widetilde{Lecture}$ is the residual of a linear projection of $\Delta Lecture$ on all other between-subject differences of control variables, represented by ΔT. The second equality holds if the other controls (T) are orthogonal to η. The condition that selection on unobservables is equal to selection on observables can be stated as

$$\frac{Cov(\Delta T'\beta_3, \Delta Lecture)}{Var(\Delta T'\beta_3)} = \frac{Cov(\Delta Lecture, \eta)}{Var(\eta)} \tag{4.7}$$

Equation (4.7) can be used to estimate the numerator of the bias of $\widehat{\beta_4}$, once we have consistent estimates for β_3. Under the assumptions that the true effect of lecture style teaching is zero and again that T is orthogonal to η, β_3 can be consistently estimated (see appendix).

The estimated bias displays the effect we would estimate even if the true effect was zero when selection on unobservables is as strong as selection on observables. In addition, we report the ratio of the estimated β_4 from equation (4.5) and the estimated

bias giving a hint of how large selection on unobservables would have to be compared to selection on observables to explain the entire estimated effect. A value higher than one indicates that selection on unobservables needs to be stronger than selection on observables to explain the entire estimate, in case of a ratio lower than one already weaker selection on unobservables than on observables suffices to explain the entire estimate.

4.5 Results

Estimates of the effect of teaching practices based on the different methods advanced in Section 4.4 are presented in Tables 4.4 and 4.5. Each regression is performed at the level of the individual student and each of the estimations also takes into account the complex data structure produced by the survey design and the multi-level nature of the explanatory variables.

Table 4.4 reports results from estimating equation (4.1) and equation (4.3). We estimate both equations separately for math and science. Columns 1 and 3 present regressions results for math and science based on equation (4.1). These regressions include a complete set of student- and family-background variables, controls for teacher and class characteristics as well as characteristics of the school. Given the purpose of this study, only estimated coefficients for the teaching practice variable of interest and selected teacher characteristics are reported.

Our key variable of interest, teaching time devoted to lecture style presentation relative to time spent on problem solving, is estimated to have a positive impact on test scores in both subjects. In math the estimate is highly significant, while the estimate in science falls short of achieving statistical significance at any common significance level.

As discussed in the previous section, these results might be confounded by between school sorting based on unobservable characteristics of students. Column 2 and 4 therefore report estimation results based on equation (4.3), which includes school fixed effects. Lecture style presentation is now highly significant in science and the point estimate significantly increased compared to column 3. The estimate in math, however, did not change, but lost its statistical significance due to increased standard errors.

To gain statistical power we pool both estimation samples and estimate equations (4.1) and (4.3) with the joint sample. This approach assumes that the effects of all right-hand side variables are identical in both subjects. Based on this estimation sample the relationship between more lecture style presentation and test scores is positive and

Table 4.4: Estimation Results OLS

	Math	Math	Science	Science	Pooled	Pooled
Lecture style teaching	0.514**	0.431	0.207	0.690**	0.380***	0.293***
	(0.205)	(0.396)	(0.132)	(0.292)	(0.111)	(0.088)
Other class activities	–0.070	–0.701	0.042	–0.461	0.022	–0.322**
	(0.298)	(0.523)	(0.161)	(0.301)	(0.140)	(0.135)
Total teaching time ($\frac{min}{week} * 10^{-3}$)	0.004	0.328	–0.329	–0.157	–0.064	0.070
	(0.657)	(2.977)	(0.559)	(0.873)	(0.412)	(0.454)
Female teacher	–0.156**	–0.171	–0.075	–0.022	–0.098**	–0.019
	(0.065)	(0.120)	(0.056)	(0.096)	(0.047)	(0.038)
Teacher younger than 30	0.125	0.074	0.116	0.256*	0.110	0.054
	(0.136)	(0.237)	(0.093)	(0.143)	(0.079)	(0.079)
Teacher's age 40-49	0.054	0.075	0.054	0.127	0.051	0.001
	(0.079)	(0.179)	(0.075)	(0.104)	(0.051)	(0.046)
Teacher older than 50	0.024	0.113	0.083	–0.023	0.034	0.016
	(0.082)	(0.181)	(0.073)	(0.128)	(0.057)	(0.059)
Teaching experience <1 years	–0.513***	–0.878***	–0.116	–0.022	–0.266**	–0.160*
	(0.155)	(0.242)	(0.167)	(0.175)	(0.105)	(0.093)
Teaching experience 1-5 years	–0.202**	–0.303*	–0.016	–0.032	–0.103	–0.077
	(0.097)	(0.178)	(0.072)	(0.101)	(0.068)	(0.069)
Teaching certificate	–0.368**	–1.229***	0.057	–0.358	–0.040	–0.132
	(0.170)	(0.318)	(0.180)	(0.383)	(0.137)	(0.161)
Constant	0.898	1.891	–0.015	2.340***	0.345	0.345
	(0.546)	(1.298)	(0.563)	(0.725)	(0.492)	(0.423)
Student background	Yes	Yes	Yes	Yes	Yes	Yes
School variables	Yes	No	Yes	No	Yes	No
School fixed effects	No	Yes	No	Yes	No	Yes
Teacher variables	Yes	Yes	Yes	Yes	Yes	Yes
Class variables	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6,310	6,310	6,310	6,310	12,620	12,620
R^2	0.313	0.518	0.337	0.501	0.309	0.476

* p<0.10, ** p<0.05, *** p<0.01

Note: Dependent variable is the standardized student test score. Lecture style measures the share of teaching time devoted to lecture style teaching instead of teaching based on problem solving. Student background includes students' gender, age in years, a dummy if born in first 6 months of the year, number of books at home, English spoken at home, migration background, household size and parental education. School variables include dummies capturing different levels of parental involvement in school activities and dummies for community size. Not reported teacher variables are teacher's major in math, science and education and years of teacher training. Class variables are class size, and a tracking indicator. Imputation indicators are included in all regressions. Standard errors clustered at the school level in parentheses.

significantly different from zero in both specifications.

The evidence presented in table 4.4 suggests a positive relation between more lecture style presentation and student achievement. However, within school selection of students based on unobservable student characteristics might drive this relationship. For instance, it is reasonable to assume that teachers adjust the use of teaching practices according to class composition and average student ability. We therefore difference out unobserved constant student traits by taking between-subject differences of test scores and all right-hand side variables as presented in equation (4.5).

Table 4.5 presents estimation results of the between-subject differences approach. We start out with a very basic specification without further controls in column 1 and successively add more controls that vary between subjects to account for subject-specific differences. In particular column 6 presents our headline results. In this specification we additionally control for all observable teacher and class characteristics presented in tables 4.1 and 4.8. It is quite astonishing to see that adding more control variables in the between-subject specification leaves our estimate for the effect of more lecture style teaching almost unchanged. Moreover, in contrast to most other control variables the share of lecture style teaching is estimated to be statistically significant throughout all specifications presented in table 4.5. Apart from lecture style teaching, additional minutes per week spent teaching the subject to the class also seem to have a substantial positive effect on student test scores. This is a very intuitive result as it simply suggests that students learning is an increasing function of total teaching time.

The specifications in columns 2 to 6 also include other class activities, i.e. the share of total time in class spent on other activities apart from lecture style presentation or problem solving, as a control variable. Other class activities include time spent on any of the other five categories in question 20. In particular, these other activities include reteaching material and accountability measures like reviewing homework and test and quizzes. Naturally, not all students understand material when it is taught for the first time so that reviewing can be beneficial for student achievement. Moreover, previous research has shown that accountability measures might have positive effects for student achievement (Wenglinsky, 2002). The overall average effect of all other class activities is, however, estimated to be small and insignificant.

The estimate of teaching time devoted to lecture style presentation relative to time spent on problem solving decreases in comparison to the regression results presented in table 4.4. This indicates that within school sorting matters for the estimation of teachers' choice variables such as the degree of lecture style teaching. Estimates for the effect of more lecture style teaching on student learning in table 4.5 range from 0.14 to

Table 4.5: Estimation Results First Difference

	1	2	3	4	5	6
Lecture style teaching	0.112*	0.118**	0.139***	0.132***	0.126**	0.127**
	(0.059)	(0.056)	(0.051)	(0.050)	(0.051)	(0.054)
Other class activities		–0.012	0.020	0.029	0.036	0.030
		(0.066)	(0.063)	(0.058)	(0.058)	(0.059)
Total teaching time ($\frac{min}{week} * 10^{-3}$)		0.438	0.498*	0.380*	0.402*	0.524**
		(0.284)	(0.265)	(0.220)	(0.235)	(0.234)
Female teacher			–0.001	–0.007	–0.011	–0.012
			(0.020)	(0.021)	(0.022)	(0.022)
Teaching certificate			0.030	0.038	0.075	0.074
			(0.063)	(0.111)	(0.122)	(0.124)
Teacher younger than 30			–0.041	–0.058*	–0.049	–0.043
			(0.038)	(0.035)	(0.037)	(0.037)
Teacher's age 40-49			0.018	0.047*	0.061**	0.052*
			(0.026)	(0.026)	(0.028)	(0.029)
Teacher older than 50			0.025	0.046*	0.050*	0.047
			(0.027)	(0.027)	(0.028)	(0.028)
Teaching experience <1 year			–0.003	–0.006	–0.005	–0.020
			(0.041)	(0.047)	(0.051)	(0.049)
Teaching experience 1-5 years			0.027	0.035	0.037	0.027
			(0.030)	(0.029)	(0.028)	(0.028)
Constant	–0.010	–0.025	0.004	0.001	–0.007	–0.021
	(0.014)	(0.016)	(0.026)	(0.024)	(0.026)	(0.025)
Class variables	No	No	Yes	Yes	Yes	Yes
Teacher variables	No	No	Yes	Yes	Yes	Yes
Limits to teach	No	No	No	Yes	Yes	Yes
Motivation	No	No	No	No	Yes	Yes
Teaching requirements	No	No	No	No	No	Yes
Observations	6,310	6,310	6,310	6,310	6,310	6,310
R^2	0.002	0.005	0.014	0.029	0.031	0.035
Bias[a)]				1.265	1.578	1.723
Ratio [b)]				0.104	0.080	0.074

* p<0.10, ** p<0.05, *** p<0.01

Note: a) Bias estimated as in equation (4.6) using condition (4.7) b) Ratio of the coefficient of lecture style presentation and the bias. Dependent variable is the within-student difference of standardized math and science test scores. All teacher variables are included as within student between-subject differences. Lecture style measures the share of teaching time devoted to lecture style teaching instead of teaching based on problem solving. Variables included in Motivation, Teacher variables and Teaching requirements are shown in table 1. Variables included in Class variables and teaching limits are shown in table A-1. Imputation indicators included in all but the first two columns. Standard errors clustered at the school level in parentheses.

0.1. Our headline estimate is reported in column 6 with an estimated size of 0.1.[8] This parameter suggests that shifting 10 p.p. of time from problem solving to teaching by giving lecture style presentations while holding the overall time devoted to these two activities constant is associated with an increase in student test scores of one percent of a standard deviation.

In turn, our results imply a negative correlation between more in-class problem solving and student achievement. This is consistent with the finding in Brewer and Goldhaber (1997) that more in-class problem solving for tenth grade students in math is related to lower test scores on a standardized test. Furthermore, the other commonly investigated teacher characteristics (e.g. gender, experience, credentials etc.) do not show significant or robust impacts on student achievement as can be seen in table 4.5. This is in line with previous findings in this literature and emphasizes the importance of the statistically significant relationship between more lecture style teaching and student achievement.

As pointed out in the previous section, estimates might still be biased due to selection of teachers into more (or less) lecture style teaching based on unobservable teacher characteristics. This concern is fostered by previous findings in the literature that emphasize the importance of unobservable teacher traits for student achievement. This raises the question: How can these results be interpreted?

The bias and ratio at the end of table 4.5 allow us to shed some light on the question of the influence of unobservables. The underlying assumption for the estimation of each bias is that selection on unobservables occurs to the same degree as selection on observables. In all columns the estimated bias is larger than the point estimate of the impact of lecture style teaching on student test scores. This is reflected in the ratios at the end of each column that are always smaller than one, indicating that selection on unobservables that is weaker than selection on observables suffices to explain the entire estimated coefficient. In our headline specification in column 6 selection on unobservables that is only 0.07 times as strong as selection on observables would explain the entire estimated coefficient given that the true effect is zero. On the one hand, we have included a great amount of control variables so that we believe that selection on unobservables is likely weaker than selection on observables. On the other hand, only very little selection on unobservables compared to the selection on observables suffices to explain the entire effect. Given this uncertainty, we refrain from

[8]It should be noted that the lecture style teaching variable is based on self-reported time-use by teachers, which is likely measured with error. If time-use is measured with classical measurement error, the estimate of the effect of lecture style teaching is biased towards zero. Hence, 0.1 might be interpreted as a lower bound for the true effect of lecture style teaching.

interpreting the results as evidence for a causal effect as the positive coefficient could also reflect selection of teachers with desirable unobserved characteristics into lecture style teaching.

This raises another question: Why would teachers with different desirable unobserved characteristics select different degrees of lecture style teaching compared to problem solving? While a reduced form approach of educational production cannot mirror the full complexity of the choices involved in the teaching process, we are, nevertheless, able to pin down the relationship between potential selection based on unobserved teacher traits and the causal effect of lecture style teaching as our estimation approach eliminated all other likely biases. If no selection based on unobservable teacher traits exists, our estimates speak for a positive effect of lecture style teaching. Our estimates might, however, be biased upwards if teachers with desirable unobserved characteristics more frequently base their instruction on lectures. Theoretically, this selection bias could be large enough to hide a true negative effect of lecture style teaching, which would imply that teachers with desirable unobserved characteristics predominately select themselves into an inferior teaching practice. This scenario, however, lacks any intuitional or theoretical support. We thus argue that this scenario is highly implausible and can be excluded, which allows a conservative interpretation of our results: We find no evidence for any detrimental effect of lecture style teaching on overall student learning.

It is important to stress that our results are limited to student achievement as measured by TIMSS test scores.[9] Moreover, our results for lecture style teaching are based on a comparison with teaching based on in-class problem solving. Loosely speaking our results suggest that on average lecture style teaching is at least not worse than teaching based on in-class problem solving. The comparison of average effects might, however, hide significant effect heterogeneities. Depending on the teacher, the students, the content taught, or other factors the one or the other teaching practice might be more effective.

Furthermore, our information on teaching practices, which is based on in-class time use reported by teachers, does not allow us to distinguish between different implementations of the teaching practices. One worry might be that especially the implementation of problem-based teaching differs substantially within our sample. Differences between the ideal and the actual implementation of interactive teaching styles that involve more problem-based teaching might also reconcile our findings with supportive evidence of

[9]To the extent that teaching methods cultivate general test-taking abilities differently or generate different long-term effects, a focus on other learning outcomes or on long-term effects might produce different results.

modern approaches to teaching (Lou et al., 1996; Machin and McNally, 2008). Thus, while a certain teaching practice may be very effective if implemented in the correct way, an empirical analysis that is based on the actual average implementation of this teaching practice might not reveal any positive effects. Our results, therefore, do not call for more lecture style teaching in general. The results rather imply that simply reducing the amount of lecture style teaching and substituting it with more in-class problem solving without concern for how this is implemented is unlikely to raise overall student achievement in math and science.

4.6 Robustness Checks

This section tests the sensitivity of the results presented in section 5 with respect to other definitions of the lecture style variable and with respect to specifications allowing for heterogeneous effects. The results of these robustness checks are presented in tables 4.6 and 4.7.

As our grouping of the response categories available to the teacher in question 20 of the 2003 teacher questionnaires in TIMSS could be criticized, we provide evidence on the effect of interest based on different approaches to construct the lecture style variable. We test four alternative definitions of the lecture style variable with corresponding estimation results presented in each of the four columns of the upper panel of table 4.6.

In column 1 time spent re-teaching and clarifying content/procedures is included in the proxy for lecture style teaching. In column 2 taking tests or quizzes is added to the problem solving category. Hence, the lecture style teaching variable in column 2 is defined in relation to the enlarged definition of teaching based on problem solving. In column 3 we decompose the variable other class activities into its elements and separately control for each category. In column 4 lecture style is defined as the share of overall time in class spent on giving lecture style presentation.

The coefficients in the upper panel of table 4.6 reveal that redefining our key variable of interest does not change the estimated impact of more lecture style teaching on student achievement. While the main purpose of this study is to analyze the teaching of new material by giving lecture style presentations rather than by letting students solve problems, it is reassuring to see that increasing the total amount of time in class devoted to lecture style presentations (in contrast to all other in-class activity) is also associated with higher student achievement.

Additionally, we present evidence for various sub-samples in the middle panel of

Table 4.6: Robustness Checks I

Other Definitions

	Def 1	Def 2	Def 3	Def 4
Lecture style teaching	0.145**	0.161**	0.134**	0.156*
	(0.060)	(0.063)	(0.054)	(0.081)
Observations	6,310	6,310	6,310	6,310
R^2	0.035	0.036	0.036	0.035

Subsamples

	Same Peers	Diff Peers	No Track	Track
Lecture style teaching	0.355***	0.082	0.116*	0.116
	(0.118)	(0.058)	(.070)	(.091)
Observations	2,205	4,105	3,529	2,292
R^2	0.096	0.046	0.065	0.071

Heterogenous effects

	Diff	Background
Lecture style teaching	0.203***	0.130*
(math)	(0.074)	(0.072)
Lecture style teaching	–0.105	–0.111
(science)	(0.073)	(0.071)
Observations	6,310	6,310
R^2	0.065	0.081

* p<0.10, ** p<0.05, *** p<0.01

Note: Dependent variables in all panels and columns are the within-student between-subject differences in standardized test scores. All teacher variables, class variables, motivation and teaching limits are included as controls. Upper panel: In Def 1 time spent re-teaching and clarifying content/procedures is added to lecture style teaching. In Def 2 taking tests or quizzes is added to problem solving. Thus, lecture style teaching in column 2 is defined in relation to the enlarged definition of teaching based on problem solving. In Def 3 the variable 'other class activities' is decomposed into its elements and these are separately included to control for each category. Def 4 takes time spent on giving lecture style presentation in relation to all other time-use categories (problem solving + other class activities). Middle panel: Separate estimation for different sub-samples: Column 1 only students with same classmates in both subjects, column 2 students with different classmates. Column 3 students who are tracked according to ability in either both or none of the two subjects, column 4 students who are tracked in at least one of the two subjects. Lower panel: Column 1 and 2 allow different coefficients in the two subjects. A negative coefficient for science variables stands for a positive association with the dependent variable. Column 2 additionally includes student background as controls. Imputation indicators are included in all estimations. Standard errors clustered at the school level in parentheses.

table 4.6. In column 1 and 2 we estimate equation (4.5) for students with the same peers in math and science and students with different peers, respectively. This distinction is motivated by the concern that the main effect might be driven by differences in the classroom composition. In the sub-sample with identical peers in both subjects our within-student between-subject identification strategy takes care of any potential peer effects. For students with the same peers in both subjects, shifting 10 p.p. of teaching time from problem solving to lecture style teaching is associated with an increase of almost 4 percent of a standard deviation. The estimate in the sub-sample with different peers decreases to 0.08 and lacks statistical significance. The results indicate that peer effects do not drive the positive coefficient of lecture style teaching.

Column 3 and 4 of the middle panel of table 4.6 report estimates separately for students in schools where either no or both subjects are tracked by ability and for students in schools where tracking on ability exists in only one of the two subjects. This distinction is motivated by the consideration that tracking on ability might induce teachers to choose different degrees of lecture style teaching. The results indicate that the positive association between more lecture style teaching and student achievement holds in both types of schools. The point estimate is the same for the two groups. In schools with differential tracking policies, however, it is not statistically significant.

In the lower panel of table 4.6 we investigate subject-specific effects. Column 1 and 2 present estimation results from estimating versions of equation (4.4), where we abandon the assumption that coefficients for each right-hand side variable are equal across subjects. As all science variables enter negatively on both sides of equation (4.4), a negative coefficient for any variable in science masks a positive relationship between the variable and the science test score. All estimates thus have the expected signs. They are not statistically significant for science, though. We thus find evidence for a stronger effect of lecture style teaching in math. A possible interpretation of the differential effects in the two subjects is that science with its natural emphasis on experimentation might just be better suited for problem solving.

So far all specifications measured time devoted to certain class activities in shares while additionally controlling for total teaching time. The specification in shares is mainly motivated by the data itself as TIMSS asks teachers to report what percentage of time in a typical week of the specific subject's lessons students spend on various activities. The data also contains information on the total time in minutes per week teachers teach math or science to the class. Thus, we can also construct a proxy for minutes per week devoted to each activity based on the information on total time per week and the shares reported by the teachers. While this constructed measure

Table 4.7: Robustness Checks II: Absolute Time Specification

Variables are measured as minutes per week spent teaching with specific teaching practice.

	(1)	(2)	(3)	(4)
Lecture style teaching	0.733*	0.751*	0.653	0.826
	(0.426)	(0.408)	(0.480)	(0.526)
Problem solving with guidance		0.184		0.174
		(0.374)		(0.538)
Problem solving w\o guidance		0.058	–0.174	
		(0.424)	(0.538)	
Other class activities	0.444	0.551*	0.362	0.536
	(0.315)	(0.297)	(0.365)	(0.461)
Total teaching time	0.150		0.236	0.062
	(0.293)		(0.374)	(0.419)

* p<0.10, ** p<0.05, *** p<0.01

Note: All teaching practice variables are rescaled as minutes per week divided by (10^3). Column (1) groups time spent on all activities but lecture style teaching and problem solving, problem solving with and without guidance as reference category. Column (2) includes all in-class time use categories, excluded is total minutes per week. Column (3) uses only problem solving with guidance as reference category. Column (4) uses only problem solving without guidance as reference category. All specifications are otherwise identical to the specification estimated in Column (6) of table 4.5. Standard errors clustered at the school level in parentheses.

presumably suffers from larger measurement error than the reported shares alone, we nevertheless estimate a version of equation (4.5) that is based on variables indicating various teaching practices measured in minutes per week as a further robustness check.

The results of this robustness check can be seen in table 4.7. All specifications measure class activities in minutes per week. Otherwise the specifications are identical to the specification estimated in column 6 of table 4.5. In column 1 of table 4.7, minutes spent per week with lecture style teaching and minutes spent on other class activities are included in the specification while minutes spent on problem solving are excluded as a reference category. One additional minute per week spent on lecture style teaching instead of problem solving is associated with an increase in test scores of 0.07 percent of a standard deviation. An increase of 23 minutes, which corresponds to a 10 p.p. shift in terms of total teaching time devoted to lecture style teaching, thus results in an increase in test scores of 1.6% of a standard deviation. This is slightly larger than our headline estimate based on shares of total teaching time. Column 2 of table 4.7 reveals that increasing the overall time devoted to lecture style teaching leads to an increase in test scores of the same magnitude as in column 1 when explicitly including the two problem solving categories and the other class activities. Columns

3 and 4 control again for total teaching time, but use either problem solving with the teachers guidance (column 3) or problem solving without guidance (column 4) as reference categories. The point estimate of lecture style teaching remains positive in both specifications, but loses significance. The point estimates for the two problem solving categories in columns 3 and 4 suggest that problem solving with guidance is slightly more effective than problem solving without guidance. The difference is, however, not statistically significant.[10]

In sum, we find positive relationships between more lecture style teaching and student achievement in all robustness analyses. The magnitude of the estimated effects varies between specifications and between sub-samples, with insignificant point estimates in some specification. Importantly, we do not find evidence for any detrimental effect of lecture style teaching in any specification.

4.7 Conclusion

Existing research on teacher quality allows two conclusions: First, there exists a large variation in teachers' ability to improve student achievement. Second, this variation cannot be explained by common, observable teacher characteristics. The results presented in this study confirm that these observable teacher characteristics have little potential for explaining the variation in student achievement. We provide, however, new evidence on a significant link between teaching practice and student achievement.

The specific teaching practice variable analyzed in this chapter is the share of teaching time devoted to lecture style presentation (in contrast to in-class problem solving). We construct this variable based on information on in-class time use provided by teachers in the 2003 wave of the Trends in International Mathematics and Science Study (TIMSS) in US schools. Exploiting between-subject variation to control for unobserved student traits and estimating a reduced form educational production function, we find that a 10 percentage point shift from problem solving to lecture style presentation results in an increase in student achievement of about one percent of a standard deviation. We further show that this result is extremely robust to definitional changes in the construction of the main variable of interest as well as to specifications allowing for heterogeneous effects.

This finding suggests that students taught by teachers, who devote more teaching

[10]Estimating corresponding specifications to columns 3 and 4 of table 4.7 based on specifications with lecture style teaching measured in percent of time as in table 4.5 reveals slightly lower estimates for the coefficient on lecture style teaching compared to our headline estimate presented in column 6 of table 4.5.

time to lecture style presentation rather than letting students solve problems on their own or with the teacher's guidance, learn more (in terms of competencies tested in the TIMSS student achievement test). This result stands in contrast to constructivist theories of learning. It is, however, in line with previous findings in the literature (Brewer and Goldhaber, 1997) showing that instruction in small groups and emphasis on problem solving lead to lower student test scores.

We emphasize, however, that our results demand a careful interpretation and need to be taken for what they are: Evidence for a positive association between more time devoted to lecture style teaching and student achievement that is neither driven by selection of particular students into schools or classes nor by selection of teachers based on various observable characteristics into a particular teaching method. However, selection based on unobservable teacher characteristics remains a worry. Following the method developed in Altonji et al. (2005), we show that only a relatively small selection based on unobservables suffices to explain the entire estimated coefficient. We thus refrain from formulating any policy conclusions that call for more lecture style teaching in general.

We are nevertheless able to draw an important conclusion about the nature of the causal effect of lecture style teaching on student achievement as we eliminated any potential biases arising from sorting of students, differences in schools and observable differences in teacher traits in our empirical approach. The existence of a sizeable negative causal effect of lecture style teaching would only be consistent with our results if teachers with favorable unobserved characteristics predominantly select themselves into an inferior teaching practice. Such a scenario, however, lacks any intuitional and theoretical support. We can thus exclude the possibility of a sizeable detrimental effect of lecture style teaching in math and science instruction on overall student achievement in US middle schools.

We believe that this result is relevant for the debate on optimizing the teaching process. Various dimensions of teaching practices have been shown to matter for student achievement. Moreover, the low-cost implementation of changes in the teaching process compared to other policy measures designed to foster student learning makes improvements in the teaching process particularly appealing. There exists, however, little consensus about what measures could improve the teaching process. Reducing the amount of traditional instruction based on lecture style teaching is typically a key candidate. Lectures are potentially connected with many disadvantages and might therefore be an inferior teaching method. National standards (NCTM, 1991; National Research Council, 1996) also advocate engaging students more in hands-on learning

activities and group work, but traditional lecture and textbook methodologies remain dominant in science and mathematics instruction in US middle schools. This raises the concern that the high share of total teaching time devoted to traditional lecture presentations has a detrimental effect on overall student learning in US middle schools. Our results, however, rather imply that, while newer teaching methods might be beneficial for student achievement if implemented in an ideal way, simply inducing teachers to shift time in class from lecture style presentations to problem solving without concern for how this is implemented contains little potential to increase student achievement. On the contrary, our results indicate that there might even be adverse effects on student learning.

Appendix

Selection on unobservables following Altonji et al. (2005)

Formally, in our application the assumption that selection on unobservables occurs to the same degree as selection on observables as imposed by Altonji, Elder and Taber (2005) can be stated as:

$$Proj(\Delta Lecture|\Delta T\beta_3, \eta) = \phi_0 + \phi_{\Delta T'\beta_3}\Delta T'\beta_3 + \phi_\eta\eta \quad (4.8)$$

$$\text{with} \qquad \phi_{\Delta T'\beta_3} = \phi_\eta \quad (4.9)$$

Where $\Delta Lecture$ captures the between subject differences in the teaching time devoted to lecture style presentation relative to problem solving and β_4 indicates its coefficient while T includes all other k control variables (teacher characteristics, time spent on other classroom activities, as well as class characteristics) and β_3 is a $kx1$ vector of coefficients. When $\Delta T'\beta_3$ is orthogonal to η the assumption (4.9) is equal to

$$\frac{Cov(\Delta T'\beta_3, \Delta Lecture)}{Var(\Delta T'\beta_3)} = \frac{Cov(\eta, \Delta Lecture)}{Var(\eta)} \quad (4.10)$$

We proceed now to answer the question how large selection on unobservables relative to selection on observabels would have to be in order to explain the entire estimate of β_4 under the assumption that the true β_4 is 0. Following Altonji et al. (2005) we regress $\Delta Lecture$ on ΔT, to get

$$\Delta Lecture = \Delta T'\delta + \widetilde{\Delta Lecture}.$$

Plugging this into our equation (4.5) yields:

$$\Delta Y = c + \Delta T'(\beta_3 + \delta \cdot \beta 4) + \widetilde{\Delta Lecture}\beta_4 + \eta \quad (4.11)$$

As $\Delta\widetilde{Lecture}$ is by construction orthogonal to ΔT the probability limit of $\widehat{\beta}_4$ can be written as

$$plim\ \widehat{\beta}_4 = \beta_4 + \frac{Cov(\Delta\widetilde{Lecture}, \eta)}{Var(\Delta\widetilde{Lecture})}$$

where

$$\frac{Cov(\Delta\widetilde{Lecture}, \eta)}{Var(\Delta\widetilde{Lecture})} = \frac{Cov(\Delta Lecture, \eta)}{Var(\Delta\widetilde{Lecture})}$$

as ΔT is orthogonal to η.

To estimate the numerator of the bias we can use the equality (4.10):

$$\frac{Cov(\Delta T'\beta_3, \Delta Lecture)}{Var(\Delta T'\beta_3)} \cdot Var(\eta).$$

For this however, we need a consistent estimate of β_3 which we obtain by estimating equation (4.11) under the assumption that $\beta_4 = 0$.

Table 4.8: Descriptive Statistics – Class Characteristics

	Math		**Science**		
	359 classes		734 classes		
Variable	Mean	SD	Mean	SD	Difference
Class variables					
Class size	23.46	6.54	24.57	7.51	-1.11**
Tracked according to ability	0.550	0.480	0.171	0.363	0.379***
Teaching limits (reference not at all/not applicable)					
Differing academic ability - a little	0.339	0.469	0.340	0.473	-0.002
Differing academic ability - some	0.330	0.466	0.321	0.466	0.008
Differing academic ability - a lot	0.204	0.399	0.174	0.378	0.031
Wide range of backgrounds - a little	0.308	0.456	0.277	0.447	0.031
Wide range of backgrounds - some	0.205	0.399	0.238	0.425	-0.032
Wide range of backgrounds - a lot	0.060	0.234	0.078	0.267	-0.018
Special need students - a little	0.309	0.457	0.328	0.469	-0.019
Special need students - some	0.147	0.350	0.184	0.387	-0.037
Special need students - a lot	0.064	0.243	0.081	0.272	-0.016
Shortage computer hardware - a little	0.140	0.343	0.236	0.423	-0.097***
Shortage computer hardware - some	0.197	0.396	0.207	0.404	-0.011
Shortage computer hardware - a lot	0.110	0.310	0.189	0.390	-0.079***
Shortage computer software - a little	0.168	0.371	0.294	0.454	-0.125***
Shortage computer software - some	0.146	0.350	0.198	0.397	-0.052
Shortage computer software - a lot	0.145	0.349	0.174	0.378	-0.029
Shortage support pc use - a little	0.181	0.380	0.217	0.411	-0.036
Shortage support pc use - some	0.148	0.351	0.185	0.387	-0.037
Shortage support pc use - a lot	0.089	0.282	0.137	0.343	-0.048*
Shortage of textbooks - a little	0.055	0.225	0.088	0.283	-0.034
Shortage of textbooks - some	0.045	0.205	0.044	0.205	0.001
Shortage of textbooks - a lot	0.011	0.103	0.083	0.275	-0.072***

Table 4.9: Descriptive Statistics – Class Characteristics (cont.)

	Math		**Science**		
	359 classes		734 classes		
Variable	Mean	SD	Mean	SD	Difference
Shortage instructional equipment - a little	0.180	0.380	0.314	0.463	-0.134***
Shortage instructional equipment - a some	0.123	0.326	0.193	0.394	-0.070**
Shortage instructional equipment - a lot	0.038	0.190	0.141	0.347	-0.103 ***
Shortage demonstrative equipment - a little	0.253	0.431	0.318	0.465	-0.065
Shortage demonstrative equipment - some	0.117	0.318	0.196	0.396	-0.080**
Shortage demonstrative equipment - a lot	0.044	0.203	0.189	0.391	-0.146***
Inadequate physical facilities - a little	0.148	0.352	0.219	0.413	-0.071*
Inadequate physical facilities - some	0.051	0.219	0.158	0.364	-0.107***
Inadequate physical facilities - a lot	0.030	0.169	0.131	0.337	-0.101***
High student teacher ratio - a little	0.230	0.417	0.292	0.454	-0.062
High student teacher ratio - some	0.132	0.335	0.204	0.402	-0.071**
High student teacher ratio - a lot	0.091	0.285	0.129	0.334	-0.038

Note: Probability weights and within school correlation are taken into account when estimating means and standard deviations. Class variables are weighted by the number of students in each class.

Bibliography

Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, 25:95–135.

Altonji, J., Elder, T., and Taber, C. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, 113(1):151–184.

Aslam, M. and Kingdon, G. (2007). What can Teachers do to Raise Pupil Achievement? The centre for the study of African economies working paper series.

Atella, V., Brindisi, F., Deb, P., and Rosati, F. (2004). Determinants of access to physician services in Italy: a latent class seemingly unrelated probit approach. *Health Economics*, 13(7):657–668.

Augurzky, B., Bauer, T. K., and Schaffner, S. (2006). Copayments in the German health system - Do they work? RWI : Discussion Papers.

Bago d'Uva, T. (2006). Latent class models for utilisation of health care. *Health Economics*, 15(4):329 – 343.

Baker, M., Stabile, M., and Deri, C. (2004). What do Self-Reported, Objective, Measures of Health Measure? *Journal of Human Resources*, 39(4):1067–1093.

Banerjee, A. V., Cole, S., Duflo, E., and Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics*, 122(3):1235–1264.

Banks, J., Crossley, T., and Goshev, S. (2007). Looking for Private Information in Self-assessed Health. HEDG Working Paper, University of York 9.

Barrow, L., Markman, L., and Rouse, C. E. (2009). Technology's edge: The educational benefits of computer-aided instruction. *American Economic Journal: Economic Policy*, 1(1):52–74.

Bonesrønning, H. (2004). Do the teachers' grading practices affect student achievement? *Education Economics*, 12(2):151–167.

Brewer, D. J. and Goldhaber, D. (1997). Why don't schools and teachers seem to matter? *The Journal of Human Resources*, 32(3):505–523.

Browne, M. and Kamiya, S. (2009). A Theory of the Demand for Underwriting. Working Paper University of Wisconsin-Madison.

Buchmueller, T., Couffinhal, A., Grignon, M., and Perronnin, M. (2004). Access to physician services: does supplemental insurance matter? Evidence from France. *Health Economics*, 13(7):669–687.

Bundorf, M., Levin, J., and Mahoney, N. (2008). Pricing and welfare in health plan choice. NBER Working Paper 14153.

Cawley, J. and Meyerhoefer, C. (2010). The Medical Care Costs of Obesity: An Instrumental Variables Approach. NBER Working Paper 16467.

Cawley, J. and Philipson, T. (1999). An empirical examination of information barriers to trade in insurance. *The American Economic Review*, 89(4):827–846.

Chatterji, P., Joo, H., and Lahiri, K. (2010). Beware of Unawareness: Racial/Ethnic Disparities in Awareness of Chronic Diseases. NBER Working Paper 16578.

Chiappori, P., Jullien, B., Salanie, B., and Salanie, F. (2006). Asymmetric information in insurance: General testable implications. *Rand Journal of Economics*, 37(4):783–798.

Cohen, A. and Siegelman, P. (2010). Testing for Adverse Selection in Insurance Markets. *The Journal of Risk and Insurance*, 77(1):39–84.

Cutler, D., Finkelstein, A., McGarry, K., and Center, L. (2008a). Preference Heterogeneity and Insurance Markets: Explaining a Puzzle of Insurance. NBER Working Paper 13746.

Cutler, D., Lincoln, B., and Zeckhauser, R. (2009). Selection Stories: Understanding Movement Across Health Plans. NBER Working Paper 15164.

Cutler, D. and Zeckhauser, R. (2000). The Anatomy of Health Insurance. In Culyer, A. and Newhouse, J., editors, *Handbook of Health Economics*, volume I, pages 563–643. Elsevier.

Cutler, J., Sorlie, P., Wolz, M., Thom, T., Fields, L., and Roccella, E. (2008b). Trends in Hypertension Prevalence, Awareness, Treatment, and Control Rates in United States Adults Between 1988-1994 and 1999-2004. *Hypertension*, 52:818–827.

De Meza, D. and Webb, D. (2001). Advantageous selection in insurance markets. *RAND Journal of Economics*, 32(2):249–262.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, 95(2):158–165.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *The Journal of Human Resources*, 42(3):528–554.

DeSalvo, K., Bloser, N., Reynolds, K., He, J., and Muntner, P. (2006). Mortality Prediction with a Single General Self-Rated Health Question - A Meta-Analysis. *Journal of General Internal Medicine*, 21(3):267–275.

Dignath, C. and Büttner, G. (2008). Components of fostering self-regulated learning among students: A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, 3:231–264.

Dixon, J. (2010). The effect of obesity on health outcomes. *Molecular and Cellular Endocrinology*, 316(2):104–108.

Doiron, D., Jones, G., and Savage, E. (2008). Healthy, wealthy and insured? The role of self-assessed health in the demand for private health insurance. *Health Economics*, 17(3):317–34.

Druss, B., Marcus, S., Olfson, M., and Pincus, H. (2002). The most expensive medical conditions in America. *Health Affairs*, 21(4):105–111.

Einav, L., Finkelstein, A., and Cullen, M. (2010a). Estimating Welfare in Insurance Markets Using Variation in Prices. *Quarterly Journal of Economics*, 125(3):877–921.

Einav, L., Finkelstein, A., and Schrimpf, P. (2010b). Optimal mandates and the welfare cost of asymmetric information: evidence from the UK annuity market. *Econometrica*, 78(3):1031–1092.

Fang, H., Keane, M., and Silverman, D. (2008). Sources of advantageous selection: Evidence from the Medigap insurance market. *Journal of Political Economy*, 116(2):303–350.

Farbmacher, H. (2010). Copayments for doctor visits and the probability of visiting a physician - evidence from a natural experiment. University of Munich Working Paper.

Finkelstein, A. and Poterba, J. (2002). Selection effects in the United Kingdom individual annuities market. *The Economic Journal*, 112(476):28–50.

Finkelstein, A. and Poterba, J. (2004). Adverse Selection in Insurance Markets: Policyholder Evidence from the UK Annuity Market. *Journal of Political Economy*, 112(1):183–208.

Finkelstein, A. and Poterba, J. (2006). Testing for Adverse Selection with Unused Observables. NBER Working Paper 12112.

GAO (2003). Private Health Insurance: Federal and State Requirements Affecting Coverage Offered by Small Business. United States General Accounting Office.

Goe, L. (2007). The Link Between Teacher Quality and Student Outcomes: A Research Synthesis. Report, National Comprehensive Center for Teacher Quality, Washington DC.

Gordon, R., Kane, T. J., and Staiger, D. O. (2006). The hamilton project: Identifying effective teachers using performance on the job. Report Washington, DC, The Brookings Institution.

Gregory, C., Blanck, H., Gillespie, C., Maynard, L., and Serdula, M. (2008). Perceived health risk of excess body weight among overweight and obese men and women: differences by sex. *Preventive Medicine*, 47(1):46–52.

Grobe, T., Dörning, H., and Schartz, F. (2010). BARMER GEK Arztreport. Schriftenreihe zur Gesundheitsanalyse.

Hanushek, E. (2002). Publicly provided education. In Auerbach, A. and Feldstein, M., editors, *Handbook of Public Economics*, volume 4, pages 2045–2141. Elsevier.

Hanushek, E. and Woessmann, L. (2009). Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation. NBER Working Paper 14633.

Haslam, D. and James, W. (2005). Obesity. *The Lancet*, 366(9492):1197–1209.

He, D. (2009). The life insurance market: Asymmetric information revisited. *Journal of Public Economics*, 93(9-10):1090–1097.

Hendel, I. and Lizzeri, A. (2003). The Role of Commitment in Dynamic Contracts: Evidence from Life Insurance. *Quarterly Journal of Economics*, 118(1):299–327.

Hurd, M. (2009). Subjective probabilities in household surveys. *Annual Review of Economics*, 1:543–562.

Hurd, M. and McGarry, K. (2002). The Predictive Validity of Subjective Probabilities of Survival. *The Economic Journal*, 112(482):966–985.

Idler, E. and Benyamini, Y. (1997). Self-rated health and mortality: a review of twenty-seven community studies. *Journal of Health and Social Behavior*, 38(1):21–37.

Kan, K. and Tsai, W. (2004). Obesity and risk knowledge. *Journal of Health Economics*, 23(5):907–934.

Kane, T. J., Rockoff, J. E., and Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York city. *Economics of Education Review*, 27(6):615–631.

Kesternich, I. and Schumacher, H. (2009). On the use of information in repeated insurance markets. Discussion paper, SFB/TR 15 Governance and the Efficiency of Economic Systems.

Khwaja, A., Silverman, D., Sloan, F., and Wang, Y. (2009). Are mature smokers misinformed? *Journal of Health Economics*, 28(2):385–397.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., and Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6):551–578.

Lo Sasso, A. and Lurie, I. (2009). Community rating and the market for private non-group health insurance. *Journal of Public Economics*, 93(1-2):264–279.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., and d'Apollonia, S. (1996). Wihtin-class grouping: A meta-analysis. *Review of Educational Research*, 66(4):423–358.

Machin, S. and McNally, S. (2008). The literacy hour. *Journal of Public Economics*, 92(5-6):1441–1462.

Manski, C. (2004). Measuring expectations. *Econometrica*, 72(5):1329–1376.

Marmot, M., Banks, J., Nazroo, J., and Blundell, R. (2009). English longitudinal study of ageing: Wave 0 (1998, 1999 and 2001) and waves 1-3 (2002-2007). 11th Edition. Colchester, Essex: UK Data Archive [distributor] SN: 5050.

Martin, M. E. (2005). *TIMSS 2003 User Guide for the International Data Base*. International Association for the Evaluation of Educational Achievement (IEA), Boston.

Matsumura, L. C., Garnier, H. E., Pascal, J., and Valdes, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment*, 8(3):207–229.

McCarthy, D. and Mitchell, O. (2010). International Adverse Selection in Life Insurance and Annuities. In Tuljapurkar, S., Chu, C., Gauthier, A., Ogawa, N., and Pool, I., editors, *Riding the Age Wave*, volume 3, chapter 6. Springer.

McTigue, K., Harris, R., Hemphill, B., Lux, L., Sutton, S., Bunton, A., and Lohr, K. (2003). Screening and interventions for obesity in adults: summary of the evidence for the US Preventive Services Task Force. *Annals of Internal Medicine*, 139(11):933–966.

Mossialos, E. and Thomson, S. (2009). Private Health Insurance in the European Union. Brussels, European Commission (Directorate General for Employment, Social Affairs and Equal Opportunities).

National Center for Health Statistics (2010). *Health, United States, 2009: With Special Feature on Medial Technology*. Hyattsville, MD.

National Research Council (1996). National science education standards. Report, Washington, DC.

NCTM (1991). Professional standards for teaching mathematics. Report, National Council of Teachers of Mathematics, Reston, VA.

Oliveira, C. (2008). Nurse Visit - ELSA user day. Presentation at ELSA user day 2008, Economic and Social Data Service.

Poterba, J. (1996). Government Intervention in the Markets for Education and Health Care: How and Why? In Fuchs, V., editor, *Individual and Social Responsibility: Child Care, Education, Medical Care, and Long-Term Care in America*, pages 277–308. University of Chicago Press.

Propper, C. (1989). An econometric analysis of the demand for private health insurance in England and Wales. *Applied Economics*, 21(6):777–792.

Risk Classification in Individually Purchased Voluntary Medical Expense Insurance (1999). The American Academy of Actuaries Issue Paper.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458.

Rothschild, M. and Stiglitz, J. (1976). Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information. *Quarterly Journal of Economics*, 90(4):629–649.

Rouse, C. E. and Krueger, A. B. (2004). Putting computerized instruction to the test: A randomized evaluation of a "scientifically based" reading program. *Economics of Education Review*, 23(4):323–338.

Rückert, I., Böcken, J., and Mielck, A. (2008). Are German patients burdened by the practice charge for physician visits('Praxisgebuehr')? A cross sectional analysis of socio-economic and health related factors. *BMC Health Services Research*, 8(232).

Schacter, J. and Thum, Y. M. (2004). Paying for high- and low-quality teaching. *Economics of Education Review*, 23(4):411–430.

Schoenbaum, M. (1997). Do smokers understand the mortality effects of smoking? Evidence from the Health and Retirement Survey. *American Journal of Public Health*, 87(5):755–759.

Schreyögg, J. and Grabka, M. M. (2010). Copayments for ambulatory care in Germany: a natural experiment using a difference-in-difference approach. *European Journal of Health Economics*, 11(3):331–341.

Scott, A. (2000). Economics of general practice. In Culyer, A. and Newhouse, J., editors, *Handbook of Health Economics*, volume 1, pages 1175–1200. Elsevier.

Seidel, T. and Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4):454–499.

Slavin, R. E., Lake, C., and Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research*, 79(2):839–911.

Smith, J. (2007). Nature and causes of trends in male diabetes prevalence, undiagnosed diabetes, and the socioeconomic status health gradient. *Proceedings of the National Academy of Sciences*, 104(33):13225–13231.

Viscusi, W. (1990). Do smokers underestimate risks? *Journal of Political Economy*, 98(6):1253–1269.

Viscusi, W. and Hakes, J. (2008). Risk beliefs and smoking behavior. *Economic Inquiry*, 46(1):49–59.

Weiss, I. R. (1997). The status of science and mathematics teaching in the United States: Comparing teacher views and classroom practice to national standards. NISE Brief 3, National Institute for Science Education, Madison: University of Wisconsin-Madison.

Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives*, 10(12):1–28.

Winkelmann, R. (2006). Reforming health care: Evidence from quantile regressions for counts. *Journal of Health Economics*, 25(1):131–145.

Wooldridge, J. (2010). *Econometric analysis of cross section and panel data*. The MIT press, 2 edition.